LUNDS UNIVERSITET
Ekonomihögskolan

Institutionen för informatik

The Impact of Generative AI in Enhancing Cybercriminal Fraudulence

A Formal Experiment on the Effectiveness of AI-Generated Frauds on Victims' Perception

Kandidatuppsats 15 hp, kurs SYSK16 i Informatik

Författare: Jakob Bjelvér
Michael Welsapar

Handledare: Markus Lahtinen

Rättande lärare: Umberto Fiaccadori
Blerim Emruli

The Impact of Generative AI in Enhancing Cybercriminal Fraudulence: A Formal Experiment on the Effectiveness of AI-Generated Frauds on Victims' Perception

SWEDISH TITLE: Inverkan av Generativ AI på Förstärkningen av Cyberkriminella Bedrägerier: Ett Formellt Experiment om Effektiviteten av AI-genererade Bedrägerier på Offrens Uppfattning

AUTHORS: Jakob Bjelvér & Michael Welsapar

PUBLISHER: Institutionen för informatik, Ekonomihögskolan, Lunds universitet

EXAMINATOR: Osama Mansour, PhD

SUBMITTED: May, 2024

DOCUMENT TYPE: Bachelor’s thesis

PAGE NUMBERS: 80

KEYWORDS:: Cybercrime, Cyberattack, Generative AI, Online fraud, Digital deception

SAMMANFATTNING: Denna avhandling utforskar inverkan av generativ AI på förstärkning av cyberkriminella aktiviteter, särskilt genom skapandet av bedrägligt innehåll och hur det påverkar offrens uppfattning. En serie formella experiment (n = 15) bedömde hur AI-genererat bedrägligt material jämför sig med mänskligt genererade bedrägerier när det gäller att påverka offer, med användning av modellen för sannolikhetsbaserad bearbetning (ELM) och relaterad kognitiv forskning. Den experimentella designen innehöll jämförelser inom grupper, med fokus på responstid, trovärdighetsskattning, samt individuell bakgrund och självbedömning. Resultaten tyder på att AI-genererat innehåll, som vanligtvis uppfattas som mindre trovärdigt än autentiska material, matchar övertygelsekraften hos mänskligt genererade bedrägerier under vissa förhållanden. Detta påvisar den oroande potentialen hos generativ AI inom cyberbrott, som kan likställas med mänskliga bedrägliga taktiker. Avhandlingen betonar behovet av innovativa cybersäkerhetsstrategier och ett informerat samhälle för att möta de utmaningar som AI-drivna cyberbedrägerier innebär. Den förespråkar fortsatt forskning och uppdaterade regleringsåtgärder för att mildra de växande hoten från AI i cyberkriminella utnyttjanden.

ABSTRACT: This thesis explores the impact of generative AI on enhancing cybercriminal activities, particularly through the creation of fraudulent content that affects victim perception. A series of formal experiments (n = 15) assessed how AI-generated fraudulent materials compare to human-made frauds in influencing victims, using the elaboration likelihood model (ELM) and related cognitive research. The experimental design involved within-group comparisons, focusing on response time, authenticity, credibility, as well as individual background and self-assessment. Findings suggest that AI-generated content, while typically perceived as less credible than authentic materials, matches the persuasiveness of human-made frauds under certain conditions. This highlights the concerning potential of generative AI in cybercrime, capable of emulating human deceptive tactics effectively. The thesis highlights the necessity for innovative cybersecurity strategies and an informed society to address the challenges posed by AI-enhanced cyber fraud. It advocates for continued research and updated regulatory measures to mitigate the evolving threats of AI in cybercriminal exploit.

Table of Contents

Figures

Tables

1 Introduction

The detection and prevention of cybercrime has traditionally relied on human caution and intervention. However, with the advent of artificial intelligence (AI), a new era of technological assistance and sophistication has emerged. Machines now mimic—and sometimes even surpass—human capabilities in various domains (Morris, 2024; Zhai et al. 2023). By using complex algorithms and supervised training on massive datasets, these systems can produce authentic-looking content comparable to what humans typically create (Hubert et al. 2020; Google, 2024).

According to Gartner (2024), generative AI is developing into a general-purpose technology capable of significantly altering society by impacting economic and social structures. The European Parliament (2020) agrees with that statement, highlighting its potential effects on various aspects of life. AI is arguably one of the most significant and evident entities on the global stage today and, as such, calls for careful consideration and management. While AI promises substantial benefits, it also poses new ethical and criminal challenges (Google, 2024; Caldwell et al. 2020; Gupta et al. 2023; Schmitt & Flechais, 2023; Ferrara, 2024). This duality makes AI a paradox, capable of driving innovation while simultaneously increasing risks if not governed with a balanced approach that considers both its extensive capabilities and its potential for misuse. Generative AI offers cybercriminals substantial benefits, notably by reducing human involvement and thus minimizing errors, but also by enabling automation, scalability, availability, and transformation (Mitchell, 2019; Caldwell et al. 2020; Google, 2024).

AI models have a high fidelity to their training data, giving their produce a close resemblance to original works which could further complicate individuals' ability to detect anomalies (Boden, 2004). Presumably the most popular generative AI models, OpenAI's GPT-series, was launched to the public in November 2022 through the release of ChatGPT. The first model in the line-up, being GPT-3, is trained on 45TB of compressed plaintext data, reported to have an estimated 100 million weekly users in just two months after its release (Brown et al. 2020; The Guardian, 2023). Numerous adaptations of this service have since surfaced, many of which rely on customizing the service to assist in unethical and malicious efforts (Falade, 2023). Although, as proven by Gupta et al. (2023) in a series of attempts to instruct the chatbot to write malicious programming code, safety measures can easily be circumvented using the original service. This implies that these technological innovations, while capable of beneficial contributions, are also accessible to those intent on engaging in criminal and unethical activities.

Cyber attackers often compromise the credibility of their attacks due to difficulty in mimicking legitimate websites and emails, news articles, phone calls, and more. However, the advancements in generative AI now provide an opportunity to refine these shortcomings, increasing the authenticity and sophistication of fraudulent content by replicating its original appeal (Google, 2024; Ferrara, 2024).

Generative AI not only improves existing tactics but also makes the creation of previously unseen and complex media like deepfakes possible, which can impersonate individuals in images, videos, and audio. This technology introduces a new dimension of cyberattacks that were previously unseen, enabling a broad set of fraudulent activities that threaten trust, democracy and can assist in information warfare (CISA, 2023; Whyte, 2020; Caldwell et al. 2020).

1.1 Problem Area

Generative AI is altering the cybercrime field, allowing malicious actors to produce realistic and deceptive materials with ease. The UK National Cyber Security Center (NCSC, 2024) states that the increasing innovation of generative AI and large language models will complicate the identification of fraudulent communications, making it challenging for individuals to discern between authentic and malicious messages. For instance, a recent case in Hong Kong involved a financial worker transferring $25 million after a video call with someone he mistakenly believed to be the company's CFO. In reality, the attack was using deepfake technology to impersonate the CFO (Chen & Magramo, 2024). Such incidents stress the growing threat posed by generative AI to cybersecurity. Combining these findings with the plethora of research that predicts the advancements of generative AI to be of an even greater concern than this present state, it is clear to say that this is an alarming topic that needs direct attention (Ferrara, 2024).

1.2 Research Question

With this background, the authors have formulated the following research questions:

1. How does AI-generated fraudulent content affect victims' perceptions of credibility and their susceptibility to fraud, compared to human-made counterparts?

2. To what extent does AI enhance cybercriminals' capabilities in creating fraud, accounting for its effectiveness in deception and the introduction of previously unseen approaches, especially considering the advancing trends in generative AI?

1.3 Purpose

The purpose of this study is to investigate how generative AI impacts cybercrime, particularly by assisting in the creation of fraudulent content. This study will assess victims' perception and judgment regarding AI-generated content inherent to cyberattacks, while also exploring the rise of new forms of threats previously unseen to the same extent as traditional methods. Through experiments measuring participants' credibility towards genuine, human-made, and AI-generated content, this research seeks to discover the extent to which generative AI complicates the identification of cyber threats by determining its effectiveness in enhancing fraud. Ultimately, this study aspires to inform policymakers and the cybersecurity community to develop and implement new and adapted strategies that can help mitigate the growing problems posed by generative AI, as well as to make individuals and organizations aware of the current state in AI-driven cybercrime.

1.4 Delimitation

This study focuses specifically on victims' perceptions of AI-generated fraudulent content compared to human-made equivalents and authentic examples. The primary types of cyberattacks examined are phishing, disinformation, and spoofing, as these are considered most susceptible to enhancement through generative AI (Google, 2024; Caldwell et al. 2020; Schmitt & Flechais, 2023). Moreover, this study prioritizes a practical, experimental approach by directly assessing victim responses to various content types, rather than attempting to predict future trends. It avoids extensive analysis of existing research findings and sidesteps the detailed assessment of surrounding mechanisms such as automation, translation, scalability, individual customization, or cultural adaptation. By concentrating on these specific areas, the research will provide fundamental insights while acknowledging the study’s limitations.

2 Literature Review

To understand the current advancements and use of generative AI within cybercrime, and the future's potential in those areas, this literature review will examine the existing research in five areas of importance (Figure 2.1). Firstly, the traditional tactics as well as the newer wave of techniques of cyberattacks will be covered to get an overview of the established and up-coming methods utilized by cybercriminals. Then, the capabilities of today's generative AI in terms of improving fraudulent material and supporting the above methods and techniques will be presented with its effects in the contemporary cybersecurity field. Furthermore, a couple of well reputable psychological theories will be presented to deeper examine what factors contribute to the victim's perception of online content and sources, to understand the role of generative AI in the malicious intent of deception. Moreover, the weaknesses and strengths of human- vs AI-generated content will be discussed relative to the above psychological foundations for deception, as well as its connection to crafting of fraudulent content. Lastly, future directions within this area will be outlined, with mentions of existing AI frameworks and their impact and use in today's AI models and services.

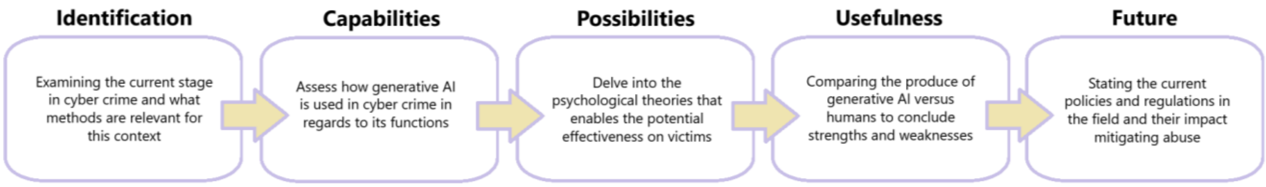


Figure 2.1: Areas of importance presented as a logical flow in literature examination.

This study will conduct an experiment to assess the implications on generative AI on cybercriminal fraudulence. Oates (2006) defines an experiment as "a strategy that investigates cause and effect relationships, seeking to prove or disprove a causal link between a factor and an observed outcome". In this case, the factor observed is the content that is exposed to victims, being authentic or fraudulent, and the outcome is the perceived credibility. An experiment is based on a hypothesis, which in turn is based on existing research and theory, and the experiment should be conducted to answer the hypothesis (Oates, 2006). Research in informatics and human-computer interaction (HCI) fields is particularly concerned with experiments, as it resides in both the human and the technical fields, making it a naturally appropriate field to conduct experiments in (Lazar et al. 2017). While correlational research is another well-fitted approach in this field, it is based on the observation of an examined phenomena, which implies that real-world measures must be taken (Field & Hole, 2003). What distinguishes this from experiments is the controlled environment that experiments are conducted in, enabling careful and direct manipulation of variables leading to better control of outcomes, although limited by the artificial setting that must be addressed as a potential bias (Field & Hole, 2003).

2.1 Conventional Cyberattacks

The following section will go over the existing cyber threat techniques and types that generative AI has the potential to enhance, as well as those predicted to have the most destructive effect on cybersecurity (Google, 2024; Schmitt & Flechais, 2023; Falade, 2023; Caldwell et al. 2020).

2.1.1 Phishing

Hummer & Byrne (2023) describe phishing as the act of via email or website persuading recipients to provide sensitive information, Personal Identifiable Information (PII), or downloading malware or ransomware that steals credentials. The report about phishing by Proofpoint (2021) indicates that the most common form of phishing is sending out the same email in mass to unknown recipients to increase the likelihood of success. Aaron et al. (2020) studied the increase of phishing websites and found that they are growing exponentially based on Google Safe Browsing’s transparency report, going from 500 000 phishing sites in 2016 to over 2 million phishing sites in 2020.

Phishing attacks will continue to be successful as long as there are humans who can be psychologically manipulated in some way (Warburton, 2020). Stalans et al. (2023) conducted a study where students (n = 236) received a phishing e-mail using the university’s phishing testing system. The email requested that they click on a link and enter their student ID to avoid having their account blocked. About half (50.8%) clicked on the link, and 81.6% of those targeted entered their PII.

2.1.2 Disinformation Operation

Disinformation operations involve the intentional creation and spread of false information to deceive and manipulate public opinion or to obscure the truth. These operations are typically state-sponsored or politically motivated and aim to influence political events, create disagreement, and damage the trust in governments and companies (Barella & Duberry, 2020). Whyte (2020) argues that disinformation operations are a threat to democracy and that cybercriminals utilize this attack in “information warfare”. These attacks can take various shapes and appear in a wide range of places, such as news articles online or posts on social media platforms (Whyte, 2020; Barella & Duberry, 2020).

2.1.3 Spoofing

The act of disguising or impersonating someone or something else in order to manipulate its victims into gaining their trust, is defined as spoofing by the US Federal Bureau of Investigation (FBI, n.d.). The objectives and motivation of spoofing varies, ranging from collecting sensitive, personal, or financial information, making payments, or spreading misinformation (FBI, n.d.). Recently, more sophisticated methods have allowed cybercriminals to create compelling and arguably more convincing depictions, with the use of generative AI (Google, 2024).

Deepfake media, enabled by generative AI to produce convincingly altered images, videos, and audio, poses notable challenges to identity verification and the integrity of information (Homeland Security, 2022; CISA, 2023). Homeland Security (2022) and a collaborative effort by the NSA, FBI, and CISA (2023) both highlight the escalating threat of deepfakes in fostering fraud, misinformation, and damaging public trust. While deepfake media does not necessarily imply spoofing, certain uses of it are inherently similar in motivation, objectives, and technique, and since the term deepfake media itself cannot be constrained as to a cyberattack, but rather a technology, this study will hereon refer to deepfake media as being a type of spoofing attack.

2.2 Capabilities of AI in Cybercrime

Having brought up the diverse shapes and techniques of cyberattacks, the question remains if generative AI possesses the capabilities to assist in the making of these. Drawing on recent scientific contributions, this study will go into detail of the dynamics between AI-powered cyber threats and the implications they have on cybersecurity countermeasures.

2.2.1 Adversarial AI in Cybercrime

Malatji and Tolah (2024) offer a comprehensive framework to understand AI's impact on cybersecurity, defining their impact on society as well as the initial motivations and defense mechanisms that can guard against these threats. In this framework, it is evident that the countermeasures needed to protect against AI-driven cyberattacks need to be present in multiple parts of society, highlighting the need for regulation and collaboration for an effective defense strategy (Malatji & Tolah, 2024). The study is centered on adversarial cyberattacks, where malicious actors exploit AI-driven cybersecurity systems to bypass defense measures, demonstrating the compound nature in using AI for enhancing cybersecurity. This study suggests concerning challenges with the introduction of AI in both attack and defense aspects, resulting in a framework that presents another dimension of cybersecurity needed to adapt to the introduction of advanced AI attack strategies, while further encouraging research for continuous innovation in this area.

Caldwell et al. (2020) covers a wide variety of cyberattacks having the potential to be AI-driven, including everything from autonomous attack drones to large scale blackmail, pinpointing their high damage potential and the challenges they pose in detection and mitigation. These findings underscore the complexity of defending against AI-powered attacks, necessitating a deep understanding of AI's capabilities. Their research results in a two-dimensional scale presenting types of AI-driven cyberattacks and their difficulty of defeat relative to the harmfulness or profitability of the crime (Figure 2.2). Interestingly in this context, is that the three types of attacks that score the highest in the combination of these dimensions are: audio/video impersonation, AI-authored fake news and tailored phishing.

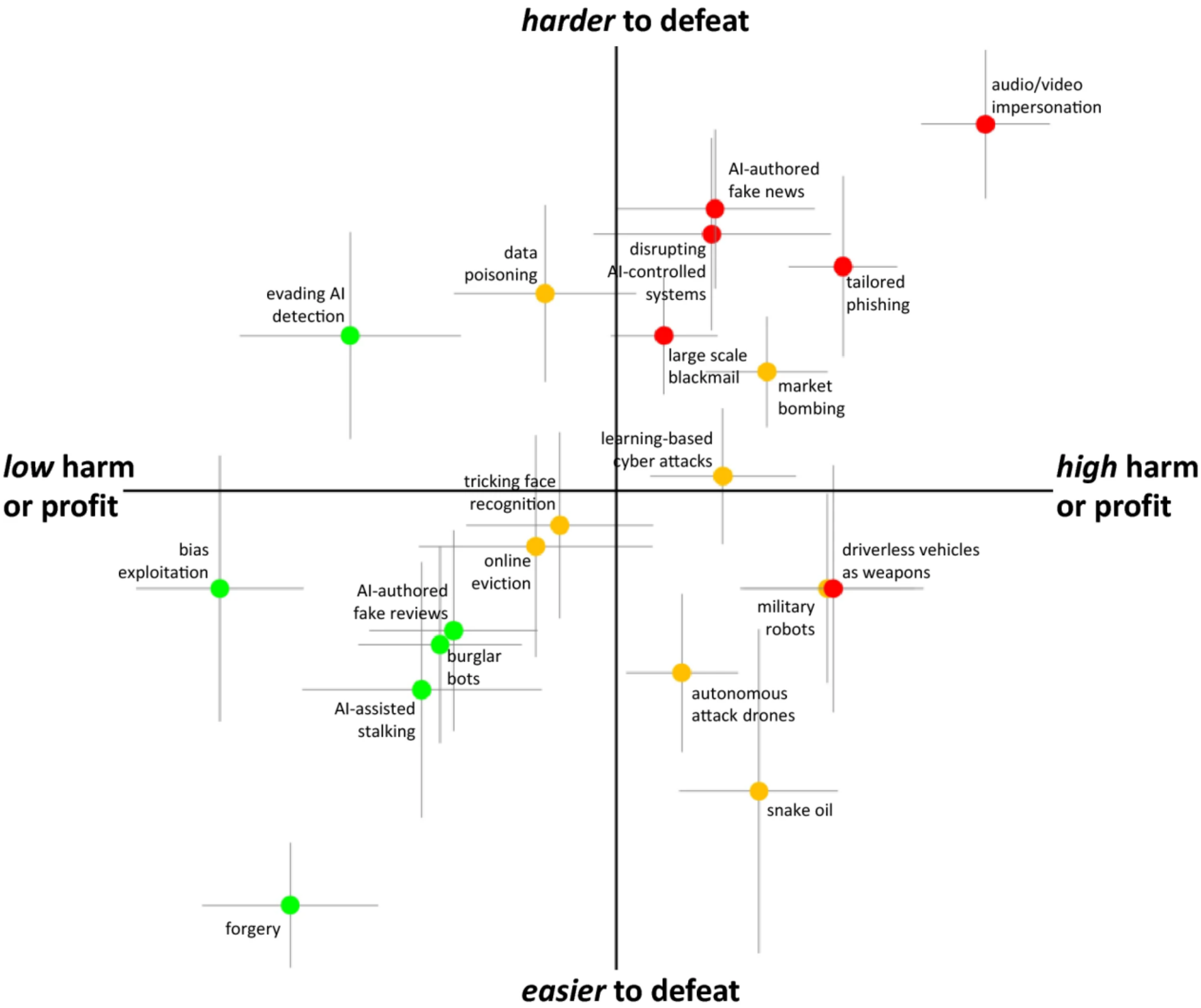


Figure 2.2: Difficulty of defeat relative to the harmfulness or profitability of the crime (Caldwell et al. 2020).

2.2.2 Generative AI and Cybersecurity

The concept of "ThreatGPT," introduced by Gupta et al. (2023), presents the wide availability and usability that generative AI presents to cybercriminals. By analyzing the ease with which security measures can be bypassed using AI, the study proves that generative AI has the potential to enable threat actors with limited resources and knowledge to utilize the power of AI to conduct sophisticated attacks, further aligning with the predictions of Google (2024). Although the research mostly focuses on how generative AI can assist in the crafting of malicious computer code, it confirms that the current security measures present in today's AI models and services are insufficient when filtering and blocking its uses for unethical, and even criminal, purposes (Gupta et al. 2023). This demonstrates generative AI's ability to be, with little effort, abused by cybercriminals to assist in the making of fraudulent content, and enables the following sections on its effectiveness in these efforts.

Schmitt and Flechais (2023) extend this discussion to the realm of Human-Computer Interaction (HCI), explaining how generative AI can strengthen social engineering attacks by generating convincing, targeted content. The research focuses on social engineering tactics, with a particular focus on different types of phishing, where the authors argue for generative AI being able to "[enhance] the effectiveness of these attacks by creating more convincing and targeted deceptive content" (Schmitt & Flechais, 2023). While this research covers the entire AI spectrum, not just generative AI, it concludes that generative AI, relevant within this study's boundaries, possesses the ability to be utilized in certain stages of developing social engineering attacks (Table 2.1).

Moreover, the study categorizes the capabilities of generative AI into four domains: text, images, voice, and video. These are essential to what Schmitt and Flechais (2023) describe as the first of four pillars, termed “Realistic Content Creation”. These domains can be directly linked to specific types of cyberattacks, such as phishing (text and websites), impersonation (voice and visual appearance), and the creation of deepfake content, which includes both images and videos. These attack types are extensively discussed and established previously (see chapter 2.1). To illustrate how these capabilities facilitate the enhancement of cyberattacks, the authors provide examples stating:

> The easiest identifiable potential of Gen AI is the creation of realistic content. A good example – especially in the context of phishing – would be website cloning. AI can rapidly clone legitimate websites and modify them subtly to deceive victims, leading to more effective phishing pages (Schmitt & Flechais, 2023, p.8).

While this study focuses specifically on social engineering attacks, covering more than just content generation, their findings strongly point to the specific capabilities that AI possess that could alter fraudulent material to be of better quality.

Table 2.1: Generative AI utilization in stages of the attack lifecycle (adapted from Schmitt & Flechais, 2023).

AI Utilization	**Stage of SE Attack Lifecycle**	**Application**	**AI Capabilities**
Attack Formulation	Goal Identification	Generating potential attack goals based on desired outcomes and vulnerabilities.	Generative AI
Preparation	Development of an Attack Vector	Crafting personalized attack vectors, like phishing emails, based on gathered information	Generative AI
Exploit Relationship	Prime the Target	Generating content aligned with the established relationship to manipulate responses	Generative AI

2.2.3 Emerging Challenges and Countermeasures

The research by Falade (2023) on "FraudGPT" and "WormGPT" investigates the underground appearance of popular AI services adapted for the purpose of misuse, revealing an increase in the accessibility of advanced unethical counterparts. In a recent report by Google (2024) the prevalence and rise of new “LLMs as a service”, offered in underground forums, for assisting in cyberattacks are predicted to increase, which is confirmed by the research of Falade (2023) with the mentions of FraudGPT and WormGPT.

Falade's study (2023) brings up two malicious adaptations of the popular, mainstream AI service ChatGPT: FraudGPT and WormGPT. Sing (2023) describes FraudGPT as a central development in the world of cyber threats: This innovative, subscription-based generative AI technology is crafted to push beyond the intended limits of ethical technology and bypass safeguards, paving the way for the creation of highly persuasive phishing communications and misleading websites. Riley (2023) found that cyber attackers have adopted WormGPT to produce persuasive, tailored emails that significantly boost the efficacy of their campaigns. Utilizing the GPTJ language model as its foundation, WormGPT is finely tuned to enhance the creation of such hostile initiatives (Riley, 2023). Originating in 2021, WormGPT offers advanced capabilities, featuring support for unlimited characters, the ability to remember conversations, and formatting tools for code. Distinct from its more ethically inclined counterparts, WormGPT is engineered with the sole purpose of facilitating illicit actions, demonstrating a particular proficiency in generating advanced and convincing phishing messages (Riley, 2023). These two offensive opposites of the apparent modified original, ChatGPT, not only demonstrates the levels of sophistication of generative AI in the cybercriminal field, but also the wide use and availability of such tools accessible in the hands of cyber criminals.

2.3 Psychological Theories Explaining Victim Perception

This section explores the cognitive mechanisms that influence individuals' perceptions of credibility and trust within digital settings, with a specific focus on AI’s capabilities in supporting deception. In essence, the advanced nature of AI-generated content would be of less importance if individuals were not put at risk of providing trust and credibility to these digital artifacts. This discussion underlines the importance of understanding the psychological base that defines how individuals evaluate the trustworthiness of information encountered in digital environments, particularly when such information is crafted with the intent to deceive with the support of generative AI technologies.

Central to this discussion are two scientific works: the Elaboration Likelihood Model (ELM) and cognitive heuristics for assessing credibility in online environments. These frameworks offer insight into the mechanisms through which individuals process information and assess its credibility, providing a foundation for understanding the impact of generative AI on cybersecurity from the perspective of potential victims.

2.3.1 Elaboration Likelihood Model (ELM)

Originally proposed by Petty and Cacioppo (1986), the ELM outlines two primary routes to persuasion: the central route and the peripheral route. The central route involves careful and thoughtful consideration of the content's argument quality, while the peripheral route relies on external cues such as the source's attractiveness or expertise. According to the ELM, individuals will take one of these two routes depending on their motivation and ability to process information. Those with higher motivation and cognitive ability are more likely to use the central route, leading to attitudes that are more enduring, resistant to change, and influential in guiding behavior. Conversely, individuals with lower motivation or limited cognitive ability are more likely to rely on peripheral cues, leading to attitudes that may be more open to change and less predictive of future behavior (Petty & Cacioppo, 1986).

The ELM framework also posits that variables like credibility, attractiveness, and repetition can influence attitudes through both routes, depending on the degree of elaboration involved. For instance, the attractiveness of a message source might serve as a peripheral cue when elaboration is low but could also impact message elaboration direction if the topic is deeper and complex (Petty & Cacioppo, 1986). Thus, the ELM provides a comprehensive framework for understanding the diverse psychological processes underpinning persuasion.

While the ELM is a widely adopted model in the psychology field, it has also been a target for critique residing in its lack of emotional incorporation, oversimplification, and assumptions about involvement (Kitchen et al. 2014; Morris et al. 2005; Stiff, 1986). Despite these critiques, Hedhli and Zourrig's (2023) study found ELM effective at predicting attitude changes, making it valuable in communication purposes.

2.3.2 Cognitive Heuristics for Assessing Credibility in Online Environments

Metzger and Flanagin (2013) posit that the credibility of online information is assessed based on cognitive heuristics. Metzger and Flanagin (2013) define credibility in the modern age as the believability of messages and that "it rests largely on perceptions of the trustworthiness and expertise of the information source as interpreted by the information receiver" (Metzger & Flanagin, 2013). Metzger and Flanagin (2013) argue that personal knowledge and experiences might influence how online credibility evaluation is processed by individuals, by allowing for bias in their cognitive heuristics that dictate decision-making and assessment of credibility. They take Internet experience as an example, where frequent online media consumption might affect metrics of credibility, such as trustworthiness and time of assessment (Metzger & Flanagin, 2013).

2.3.3 Application to Cyberfraud

Integrating the ELM and the works of Metzger and Flanagin (2013) with research on cyberfraud, these reveal how individuals process and evaluate online information, directly linking to the strategies employed by cyberattackers. Cybercriminals often exploit peripheral cues like urgency to manipulate individuals into taking impulsive actions, such as clicking malicious links or providing sensitive information (Stalans, 2023). These cues cater to individuals who, due to low motivation or cognitive ability, are more likely to take the peripheral route when processing deceptive content and therefore relying on memory shortcuts to make quick and automatic responses (Petty & Cacioppo, 1986; Metzger and Flanagin, 2013). Other times, cyberattacks appear as more engaging, in-depth content affecting victim's opinions and emotions (Whyte, 2020; Barella & Duberry, 2020). According to the ELM, these attacks would better utilize argumentative qualities and reasoning to deceive individuals (Petty & Cacioppo, 1986). Connecting psychological theory to cyberfraud, Stalans (2023) relates to the ELM by acknowledging the "dual-processing" of credibility: the heuristic approach and the systematic approach. Stalans (2023) also points out that individuals use different approaches depending on the situation:

> People often have time pressure and cognitive overload and reserve their cognitive effort for difficult tasks. Reading and responding to e-mails typically is not considered a challenging task (Stalans, 2023, p.4).

This would suggest that different cyberattacks, because of their apparent form and media, would beneficially utilize different content cues in order to effectively deceive victims. With this argument, phishing emails would, interpreting Stalans (2023), be more effective strengthening its attractiveness and urgency because of its light and short format. On the contrary, would fake news articles be more effective focusing on argument quality and reasoning logic as they are typically presented in a longer and deeper format. One aspect that all authors agree on is that the presence of urgency or the limited time in which individuals must make decisions highly affect the way that they assess credibility in, which in turn could define the strengths and weaknesses of fraud by its content attributes.

Regarding the effect of individual differences in credibility assessment, the above works are united in concluding that these play a role in the perception of information (Petty & Cacioppo, 1986; Metzger & Flanagin, 2013; Stalans, 2023). Experience and knowledge, as well as self-control and emotional character are factors contributing to assessment. Although, due to the varying effects of victimization of cyberfraud, like misinterpreting the reason why they were victimized or emotionally associating counteractions in the future, most prior victims have a limited use of learning from their victimization (Stalans, 2023).

2.4 Comparing Human and AI-Generated Content

2.4.1 Human-made Content

The human mind is able to generate a wide range of materials in creative ways by utilizing the complex nature of the human brain. In literature, for example, humans are able to compose poems, books, and essays that reflect on the individual’s unique experiences and perspectives (Culler, 2000). Likewise, artists possess the capability to produce visual artworks that express creativity and imagination (Boden, 2004).

This makes the human-made content filled with subjectivity, emotions, and cultural context, as a result of thinking and expressing thoughts and beliefs (Boden, 2004). While this paves the way for bias and personal values affecting logic and reason, it simultaneously makes the content authentic and original which resonates with audiences on a passionate level (Sloboda, 2001).

2.4.2 AI-Generated Content

In contrast, AI-generated content is produced by machine learning algorithms and computational models trained on large datasets. In recent years, AI systems have demonstrated remarkable capabilities in generating text, images, music, and other forms of creative content (Goodfellow et al. 2016). For example, AI algorithms can generate realistic images, compose musical compositions, write programming code, and even author articles or stories (Hubert et al. 2024).

AI-generated content relies on statistical patterns and mathematical algorithms to generate outputs that mimic human-produced content (Goodfellow et al. 2016). These algorithms analyze vast amounts of data to identify underlying patterns and structures, which are then used to generate new content (Hubert et al. 2024). While AI-generated content may lack the emotional depth and subjective interpretation of human-made content, it often exhibits a high degree of realism and fidelity to the training data (Boden, 2004).

2.4.3 Comparing Generated Content

The comparison between human-made and AI-generated content reveals distinct differences in terms of creativity, authenticity, and emotional resonance. Human-made content reflects the unique perspectives and creative intuition of individual creators, capturing the richness and complexity of human experience (Culler, 2000). In contrast, AI-generated content relies on computational algorithms to generate outputs based on statistical patterns and data-driven models (Goodfellow et al. 2016). While AI-generated content may achieve impressive levels of realism and accuracy, it may lack the depth, nuance, and emotional resonance of human-made content (Sloboda, 2001).

In a study performed by Májovský et al. (2023), the authors generated scientific studies with the help of a widely available generative AI model, GPT-3, and compared these with authentic ones written by humans. They concluded that the results "[...] look sophisticated and seemingly flawless, [but] expert readers may identify semantic inaccuracies and errors upon closer inspection". This demonstrates that the outcomes of generative AI may be factually incorrect and may fail to complete complex tasks, potentially deceiving professionals in the respective fields. Nevertheless, it still conforms to the peripheral route described by Petty and Cacioppo (1986) in the Elaboration Likelihood Model (ELM), by possessing strengths that initially appear correct, even without further elaboration. In practice, this means that generative AI could enable cyber criminals to craft fraudulent material to deceive novice, or even intermediate victims, which is also emphasized in the works of Gupta et al. (2023) and Falade (2023).

Further confirmation of these observations can be found in the research conducted by Zhai et al. (2024) and Morris (2023). Zhai et al. (2024) conducted a study comparing the efficacy of generative AI versus human-made content in students' science assignments. Their findings consistently demonstrated that generative AI outperformed human-made content in terms of quality and effectiveness. While this study primarily focused on educational assignments, its implications extend to the broader context of content generation, suggesting that generative AI possesses the capacity to produce high-quality text.

Morris (2023) examined the perceptions of scientists regarding the potential applications of generative AI across various domains. Through interviews with scientists, Morris revealed a consensus among respondents regarding the utility of generative AI in augmenting scientific work. Scientists expressed confidence in the ability of generative AI to support a wide range of tasks, indicating its potential to enhance efficiency and productivity in scientific efforts (Morris, 2023).

In the context of this study, the realism and precision inherent in AI technology makes it a potentially attractive instrument for cybercriminal activities. Enhancing and preventing the previously discussed common pitfalls within this unethical domain, the utilization of AI can therefore be argued to potentially have the capacity to improve the quality of cyber threats.

This is facilitated by AI's diverse capacity to, amongst many capabilities, enhance text processing and faithfully replicate authentic content for malicious intent, as suggested by the works of Zhai et al. (2024), Morris (2024) and Hubert et. al (2024). This research further supports the works of Caldwell et al. (2020), Gupta et al. (2023), Schmitt and Flechais (2023) and Falade (2023) claiming the benefits of using generative AI for malicious purposes. In turn, these results are especially putting emphasis on Caldwell et al.'s (2020) determination of the previously mentioned most harmful and difficult to defeat cyberattacks present today. Although, there are certain areas where AI is incapable of replicating the human mind, for example in the cultural or personal expressions, which one could argue would make it more susceptible to being recognized as being counterfeit or inhumanly crafted (Sloboda, 2001).

2.5 Future Directions for Cybersecurity Protocols

Ferrara (2024) investigated the future implications of generative AI for cybersecurity strategies, stressing the rapid progression of AI technologies and their integration into cybersecurity solutions. The discussion draws attention to the imperative of maintaining a step ahead of cybercriminals, who are swiftly adopting AI to engineer more effective fraudulent content. Ferrara (2024) advocates for ongoing innovation in AI-centric cybersecurity solutions as a means to safeguard against the increasingly complex attacks enabled by AI advancements.

The study underscores the urgent need for security measures as a result of AI being in the hands of cybercriminals. This requirement is magnified by the observation from Malatji and Tolah (2024) that highlight the comprehensive ways in which cybersecurity needs to evolve by innovation and collaboration to adequately protect against the misuse of this advancing technology. Thus, a coordinated approach that integrates the ethical guidelines, accountability frameworks, and standard-setting initiatives from leading global organizations is crucial.

2.5.1 Established Standards and Principles

UNESCO's (2024) global standard on AI ethics recognizes the importance of protecting human rights and dignity in the AI era. UNESCO's framework, emphasizing transparency, fairness, and human oversight, provides a foundational ethical blueprint for AI development. It reflects a global consensus on the need for AI technologies to be underpinned by core human values, ensuring they contribute positively to society and the environment.

Simultaneously, the OECD's (2019) efforts to foster international collaboration on AI governance, through principles of algorithmic accountability and the establishment of expert networks, stresses the significance of global interoperability and shared ethical standards. These principles not only aim to facilitate innovation and trade but also ensure that AI systems are developed and deployed in a manner that respects ethical norms and societal expectations.

Furthermore, ISO's (n.d.) commitment to setting international standards for responsible AI and ethics—focusing on transparency, data protection, and user privacy—offers a tangible framework for assessing and guiding AI development. These standards represent a crucial tool for stakeholders across the AI domain, from developers to policymakers, in ensuring that AI technologies adhere to established best practices and ethical considerations.

2.5.2 Integration into AI Development

Ferrara (2024) and Malatji and Tolah (2024) argue for a multi-faceted approach to enhancing and securing AI technologies, one that seamlessly integrates the ethical, regulatory, and standard-setting initiatives from UNESCO (2024), OECD (2019), and ISO (n.d.). By embedding these global frameworks, together with the perspectives of Ferrara (2024), into the fabric of AI development and deployment, stakeholders can address the vulnerabilities identified by Malatji & Tolah (2024), mitigating the risks of misuse and exploitation as mentioned by Gupta et al. (2024).

Ferrara (2024) identifies multiple recommendations for supporting authenticity and source credibility, such as watermarks, certificates and blockchain technology, and thereby improving current regulations and safety measures, but at the same time is it evident that very few of these are present into the frameworks that govern the production and servicing of AI.

2.6 Summarizing Existing Research

The literature review has provided an in-depth exploration of various aspects surrounding the intersection of generative AI and cybercrime. By examining existing research, this review has shed light on the evolving cybersecurity field, the capabilities of AI in facilitating cyber threats, and the psychological frameworks influencing victim perception. Additionally, it has compared human-made and AI-generated content while outlining future directions for enhancing cybersecurity protocols.

Conventional cyberattacks, such as phishing, disinformation operations and spoofing, remain prevalent and pose significant challenges to digital security. The discussion highlighted the adaptation of these traditional tactics to leverage generative AI, as well as the emerging techniques that follow the development of this accelerating technology.

The review highlighted the significant role of AI in cybercrime, illustrating how AI technologies can be exploited by cybercriminals to craft sophisticated attacks. Malatji and Tolah (2024) provide a comprehensive framework showcasing the alterations that cybersecurity needs to undergo to keep up with the evolving capabilities of AI-driven attacks. These adversarial AI tools pose challenges for cybersecurity professionals, necessitating ongoing research and development of AI-driven defensive mechanisms. Caldwell et al. (2020) further accentuate the urgency for innovative cybersecurity approaches to mitigate the risks posed by AI-powered attacks, and it can be argued that the threats found the most serious are also exemplified as having the greatest potential to be assisted and enhanced by generative AI as per Schmitt & Flechais (2023). Additionally, Falade's (2023) research on "FraudGPT" and "WormGPT" highlights the sophistication of modified versions of mainstream AI services intended for malicious use, and its availability and wide-spread adoption among cybercriminals. Ultimately, it can be said that AI is difficult and complex to protect against, while its harm relative to its ease of use are concerning seen to the availability and suitability in generating these attacks.

Furthermore, the literature explored established frameworks for understanding victim perception within digital environments. The Elaboration Likelihood Model and cognitive heuristics in the perception of digital content offer valuable insights into how individuals process and evaluate information online, particularly in the context of AI-generated deceptive content Integrating psychological foundations with research on AI-generated content highlights the various possibilities for AI to replicate authentic cues of credibility that can deceive humans (Petty & Cacioppo, 1986; Metzger & Flanagin, 2013).

A comparative analysis between human-made and AI-generated material revealed distinct differences in creativity, character, and emotional resonance. While AI-generated content exhibits impressive levels of realism, it may lack the depth and nuance characteristic of human-made content. However, recent studies suggest that AI-generated content can outperform human-made content in certain tasks, especially in replication, underscoring its potential to be used maliciously in cybercrime with certain implications (Májovský et al. 2023; Morris, 2023; Boden, 2004).

Finally, the literature review outlined future directions for enhancing cybersecurity protocols, emphasizing the importance of ethical AI development and responsible deployment. By continuously innovating and integrating adapted global ethical frameworks into AI development and governance, stakeholders can mitigate the risks of AI misuse and exploitation, ensuring that AI technologies serve as a force for good in society (Ferrara, 2024; Malatji & Tolah, 2024; UNESCO, 2024).

In conclusion, the literature review provides a comprehensive overview of the current state of research on generative AI and cybercrime. Although much of the examined research is directed in the futuristic and theoretical manner with little to no research on the actual effects on victim's perception and therefore the success rates of AI-generated cyberattacks, which leaves the question open on what impact the proven capabilities and adoption of generative AI has on in the cybercriminal landscape in terms of documented measures.

Table 2.2: Compilation of references by area of importance.

Factors	**Keywords and explanations**	**References**
Advanced AI in assisting conventional cyberattacks and providing new and innovative methods	• Cyberattacks, vulnerabilities Traditional cyberattacks persist and are adapting with generative AI, necessitating robust defense mechanisms.	Hummer & Byrne (2023), Whyte (2020), Barella & Duberry (2020), FBI (n.d.), CISA (2023)
Insufficient safeguards in AI services and models; Misuse of capabilities of generative AI; Cybersecurity challenges posed by AI	• Generative AI, cybercrime AI technologies exploited by cybercriminals to craft sophisticated attacks.	Malatji & Tolah (2024), Caldwell et al. (2020), Schmitt & Flechais (2023), Falade (2023), Google (2024)
Victims' perception on cyberattacks; Understanding of the psychological level of deception	• Victim psychology, theories, persuasive factors Elaboration Likelihood Model and cognitive heuristics inform understanding of how individuals process AI-	Petty & Cacioppo (1986), Metzger & Flanagin (2013), Stalans (2021)

	generated deceptive content.	
Understanding the strengths of AI as a complement or replacement; Comparison of human-made content to determine AI's efficiency	• Content, authenticity, realism AI-generated content exhibits realism but may lack nuance; has potential for outperforming human-made content in certain tasks.	Májovský et al. (2023), Morris (2023), Boden (2004), Hubert et al. (2024), Zhai et al. (2024), Goodfellow et al. (2016), Sloboda (2001), Culler (2000)
Availability of information to support decision-making and adaptation of conventional cybersecurity protocols	• Ethics, responsibility Emphasis on ethical AI development and responsible deployment to mitigate risks of AI misuse; Innovation and integration of global ethical frameworks.	Ferrara (2024), UNESCO (2024), OECD (2019), ISO (n.d.),

3 Methodology

After having developed a comprehensive literature review and a research question, the decision of what research method to use remains and will be covered in this chapter.

This study chose to perform a formal experiment to assess the impact on victims' perception of AI-generated content. Due to the nature of this research, partly residing in the psychological field and partly in the technological field, an experiment offers the possibility to observe a causal relationship while allowing for careful control over variables maintained. As control measures, authentic content will be provided to conclude this impact. The three variants of content in this study, authentic, AI-generated, and human-made, together with the presence of urgency, make up for the independent variables of the study, while the dependent variables to be measured are response time, perceived credibility, as well as comments and potential questions.

In practice, participants in this experiment will be exposed to materials that are either authentic, AI-generated, or human-made, in either an urgent or non-urgent setting, in order to measure the credibility and to assess persuasive factors of AI-generated content. Preceding and succeeding the exposure, participants will answer questions about their personal experiences and knowledge relevant to the context, as well as provide a self-assessment of their perceived performance in correctly assessing materials as more or less authentic.

To achieve consistency and to eliminate potential biases or errors, the within-group experimental design is utilized. This includes randomization of the variations of the content exposed to participants, while ensuring the possibility of data analysis that is meaningful when comparing the efficiency of the content on the victims' perception. The procedure that follows this design is made to follow established ethical standards to ensure participant safety, while also mitigating the risks of the artificial setting of the laboratory nature in order to mimic the real-world as much as possible. As for the participant selection, a combination of techniques and requirements are utilized to achieve representativeness and to increase the reliability and external validity of the study. Considering all of the above, relevant limitations are raised and their importance is highlighted when interpreting the final results.

It should be noted that researchers often use the term "experiment" broadly to describe their research approach. However, in the context of this study, the precise term for this study's method is a "formal experiment" (Oates, 2006). For simplicity, the authors will be using these terms interchangeably going forward, as they both refer to the same research method.

3.1 Justification

Considering the various aspects encompassed within this subject, including advancements in generative AI and the dynamics of human-computer interaction in digital threats, this study's authors have opted to conduct formal experiments in a controlled environment to shed light on the cause-and-effect relationships at play. The experimental method permits researchers to manipulate diverse factors and examine their interplay, a critical aspect of this investigation (Oates, 2006). Here, the focus shifts from AI-generated fraudulent content to conventional, human-made fraudulent content, with a particular interest in assessing whether individuals fall victim to these attempts.

Given the aim to observe phenomena, the experimental method offers an advantageous framework by providing control over the research environment. While qualitative research offers valuable insights into individual experiences, it lacks the capacity to quantify and adjust parameters, particularly in exploring human-technology interactions (Oates, 2006). On the other hand, quantitative research excels in gathering numerical data but falls short in adequately addressing the multifaceted nature of this topic, which intersects technology and society. Hence, the experimental method emerges as the most suitable approach for closely examining cause-and-effect relationships (Oates, 2006; Shadish et al. 2002).

The main benefit of the experimental approach, given the circumstances and aims of this study, is the fact that it is the only research strategy that can effectively determine a causal relationship (Oates, 2006; Shadish et al. 2002). Another benefit worth mentioning is the fact that the laboratory setting used in this study provides cost-effective and reliable conditions, where the authors can continue to be present in their normal workplace (Oates, 2006; Tichy, 1998). Although the latter also means that the observations will occur outside of the real-world setting, which makes the results less credible when documented in an artificial environment (Oates, 2006; Jarvenpaa et al. 1985). This is a crucial consideration to have in mind, as observations are meant to reflect natural behaviors.

3.2 Hypothesis

For this study, two sets of hypotheses are formulated to achieve the objectives and to address the research question effectively. The primary hypothesis directly responds to the research question by employing both a null hypothesis and an alternative hypothesis. These are designed to test for a causal relationship between the identified variables or to establish the absence of such a relationship. The structure of these hypotheses is informed by Rosenthal and Rosnow's (2008) framework.

Given that this experiment involves two independent variables—namely, the type of generation method and the prevalence of urgency—it necessitates two corresponding sets of hypotheses. The secondary hypothesis explores which factor significantly influences the primary factor, with the potential outcomes of these hypotheses being either jointly or separately true or false. As discussed in the next chapter, the prevalence of urgency is the independent variable that will be used to determine what route of persuasion the type of content source will take.

∆ Hypothesis 1a: AI-generated fraudulent content is perceived as more credible than human-made fraudulent content.

∆ Hypothesis 1b: Participants are less likely to be deceived by AI-generated fraudulent content compared to human-made fraudulent content.

∆ Hypothesis 2a: The attractiveness of AI-generated fraudulent content enhances its perceived initial credibility through urgency cues being prevalent.

∆ Hypothesis 2b: AI-generated fraudulent content is perceived as more credible due to its argument quality following careful consideration without any urgency prevalent.

3.3 Experimental Design

First and foremost, the need for iterations in the experimental design needs to be addressed, as it greatly affects the reliability and research outcomes (Oates, 2006). Therefore, the experiment needs to be repeated to ensure stability and reproducibility of results across different trials. What determines the repetitiveness and number of iterations is the differential in the measures taken—if the results differ more than explainable, the experiment needs to be conducted again.

This study will utilize the within-group experimental design, as described by Field and Hole (2003) and Lazar et al. (2017). This approach relies on the repeated measures method, where the same participants are exposed to various manipulations of the independent variables. In this study, each participant will be exposed to all combinations and alterations of the materials being the independent variable in a randomized order. The randomization inherent to this design aims to eliminate biases that may arise from individual differences, thereby enhancing both the reliability and the internal validity of the findings (Lazar et al. 2017; Oates, 2006).

3.3.1 Independent Variables

The independent variables used in this study are the content source, and will exist in variations of either authentic, human-made fraudulent or AI-generated fraudulent, as well as the prevalence of urgency, that is either urgent or non-urgent. When conducting experiments, a control variable is used to measure the effects that occur due to the artificial setting of the experimental procedure (more famously known as the "placebo effect" in clinical studies) (Lazar et al. 2006). In this study, the control variable is the authentic version, which will account for the bias introduced by participants' awareness of the fraudulent presence. What remains are the human-made and AI-generated contents, that will allow for answering the first hypothesis (see chapter 3.2) by measuring the differences in perceived credibility. The other independent variable, being the presence of urgency, is related to the second hypothesis and will ultimately be used to conclude the strengths and weaknesses of AI-generated content by determining its effective route to persuasion, aligning with the psychological foundations relative to victimization of fraud (see chapter 2.3). In the case of urgency, this is done by motivating participants to use their intuition and therefore their cognitive heuristics, thus limiting participants’ ability to elaborate on the credibility of the content presented. For the non-urgent setting, participants are allowed to deeply examine the materials using a systematic approach and by judging the content’s argumentative qualities.

An important deviation to note is the variation in the levels of independent variables, specifically the number of different content sources per material due to the lack of quality counterparts to certain materials (see chapter 3.10).

3.3.2 Dependent Variables

These dependent variables are directly influenced by the independent variables and are used to measure the actual effects (Lazar et al. 2006; Field & Hole, 2003). These variables defined the outcomes of the experiment and will be used to propose the causal relationship central to the study (Field & Hole, 2003). For every exposure of a variation of the content source and urgency, a credibility rating on a scale of 1-5 together with the response time and additional comments and questions will be measured, which makes up for the dependent variables

(Table 3.2). The credibility rating is essential for ultimately determining the perceived authenticity of the content exposed, while the response time is used to control the experiment's efficacy in mediating the urgent setting to the participants. In other words, if a participant is instructed to answer quickly based on intuition, but simultaneously takes longer to answer than expected of the urgent setting, response time will be used to normalize this deviation by allowing for the determination of the effect urgency had on the participant's response. Lastly are comments and questions noted during the exposure, in order to capture participant's reasoning and to ultimately be given insights into the thought process that went along with the assessment.

3.4 Materials

Serving as the central matter in the experiment, the materials used for this study represent the alterations of the variables (see chapter 3.3) in the experimental design and are presented in Table 3.1. Essentially, the materials depict the cyberattacks examined in this study that are relevant for investigating the implications of AI on them.

3.4.1 Collecting Human-made Frauds

The process began by identifying real-world examples of frauds that were clearly human-made. To ensure authenticity, only instances of fraud from before 2021 were considered—predating the release of the first widely accessible AI model, GPT-3, introduced by OpenAI in 2022. Although AI technologies were available before this, the cutoff provides a sufficient demarcation of when AI-generated cyberattacks started to surface. Each identified fraud was paired with its authentic counterpart to establish a baseline for comparison and minimize any bias stemming from the differing contents. Furthermore, to assure sufficient quality of the collected material, solely fraudulent content that has had an impact and succeeded in deceiving people was collected. While this is a vague criterion to estimate, the authors looked at contextual cues, such as the existence of contradictory news articles exposing the content as fake, in order to assess its quality as a cyberattack. It is important to note that some materials were artificially aided by the authors to achieve a certain quality standard or in order to eliminate bias, those alterations are further detailed in relation to the corresponding materials in Appendix A.

3.4.2 Crafting and Gathering of AI-Generated Frauds

Following the identification of human-made frauds, AI-generated equivalents were produced using advanced generative AI techniques, referred to as "jailbreak" methods (Gupta et al. 2023). These methods are designed to circumvent traditional security protocols that are typically inadequate at filtering sophisticated AI outputs. Although, most of the time these methods were not needed, as the AI services would complete the request without the need for "jailbreaking". The generation process involved iteratively refining the input prompts to the generative AI service (as seen in a compiled format and correspondingly to each content in Appendix A), until the output's quality was maximized. The AI was either prompted with generating attractive content with a sense of urgency or prompted with crafting reasoning and convincing content with a focus on argument quality—based on the predefined prevalence of urgency. In other cases, where the AI-generated material was already existing, the material was sourced and confirmed to be generated using generative AI.

To summarize the materials used, the authors have gathered three versions, representing the levels of independent variables, for each examined type of cyberattack; AI-generated fraudulent, human-made fraudulent, and authentic material.

Table 3.1: Materials used in the experiment that are exposed to the participants.

Attack type	AI functionality	(ID) Description and (urgency)	Content sources
Phishing (websites)	Code generation	(7) E-commerce login page (non-urgent) (2) Delivery tracking page (urgent)	Human, AI, Authentic
Phishing (email)	Text generation, Code generation	(8) Streaming service account suspension notice (urgent) (1) Bank promotional email (non-urgent)	Human, AI, Authentic
Disinformation operation	Audio generation[1], Text generation	(3) Government official delivering a speech to the nation (urgent) (4) News article themed by political tensions (non-urgent)	Human[2], AI, Authentic
Spoofing	Video generation, Audio generation, Text generation	(5) CEO talking about a newly launched service (non-urgent) (6) Former US President in video interview (urgent)	AI, Authentic

[1] *Not applicable for News article themed by political tensions.*

[2] *Not applicable for Government official delivering a speech to the nation.*

3.5 Participant Selection

The choice of who to include in a study is crucial in ensuring that the participant pool is representative of the target population, which in this case is the general population (Field & Hole, 2003). The following section will deal with the factors that contribute to mitigating risks of internal invalidity by minimizing the individual's effect on the independent variables. It is essential that the measures of the dependent variables are not influenced by factors other than the independent variables—thus, all participants need to have as close a relation as possible to the independent variable (Lazar et al. 2017). Below are the two factors needed to vary in order to recruit participants who can represent the general population, based on the argument that experience with relevant subject might affect credibility ratings, as per Metzger and Flanagin (2013):

1. IT knowledge (which infers the use and experience of AI) – could affect the ability to recognize AI-generated material.
2. Previous contact with cyberattacks/cybersecurity experience – could affect the skills needed to identify fraud in human-made fraudulent content.

As many participants were recruited as possible to minimize the threat to the study’s ability to represent the population as a whole (Field & Hole, 2003). While this is a limitation further discussed below (see chapter 3.10), the number of participants is also not the only factor contributing to generalizability. To prove statistical significance, the results must vary only by what’s explainable; therefore, factors such as participant history and demographics, as previously discussed, are equally effective in ensuring population representation.

As for the representation of the target group, this study will utilize the non-probabilistic sampling method. It is chosen since achieving representativeness is a complex task when targeting the population as a whole, as it is simply not feasible to derive a sample from all humans on the planet (Oates, 2006). While it is still important to achieve this representativeness, the authors have chosen to deploy the following sampling techniques: Purposive sampling and Convenience sampling. The former being hand-picked representatives that are known to interact with IT in their daily life, therefore chosen by the fact that they will likely produce valuable data to meet the purpose of this research (Oates, 2006). The latter being participants that are easy for the authors to reach and willing to participate, as acquaintances and fellow students (Oates, 2006). While it is important to note that the latter should not make up most of the representatives, it can still be argued as a valuable technique for the purpose and in the context of this study (Oates, 2006).

With this background, the authors contacted organizations in both the public and private sectors, requesting the participation of employees with varied roles and duties, as well as students from a wide range of fields of study. The rationale behind this was to eliminate the potential effect that the public or private sector has on their investment in cybersecurity training, and to recruit mixed roles in order to reduce the risks of a specific skill impacting IT experience and therefore potential experience with generative AI. The choice to include students is motivated by their typically younger age, as well as their differences in background, lifestyle, and experiences from working professionals (Field & Hole, 2003). The attempt to vary participants' backgrounds will directly affect the study’s internal validity by mitigating the effects of participants' individual histories on the measures of the experiment (Field & Hole, 2003).

The above effort resulted in two organizations willing to participate, with 3 from the private sector, and 5 from the public sector, as well as 7 students from varying institutions, compiling to a total of 15 participants.

3.6 Procedure

In order to standardize the approach to this experiment to ensure reliable and measurable results, it is of utmost importance to determine and follow a detailed procedure (Field & Hole, 2003; Lazar et al. 2017). In this section, the authors will go over the complete procedure and set of activities and tasks, step-by-step, that makes up for the central event of the experiment.

3.6.1 Introductory Communication

The introductory phase of the experiment serves as a crucial foundation for participant engagement and comprehension (Field & Hole, 2003; Jarvenpaa et al. 1985). Participants will first receive a briefing about the study's objectives, the nature of the content they will evaluate, and the process of the experiment. This briefing aims to prepare participants for the tasks they will perform and ensure their comfort and understanding (Appendix B).

3.6.2 Consent

The acquisition of informed consent of participants is of great importance in ethical research practice (Lazar et al. 2017). Participants are provided with clear, written consent forms detailing the study's purpose, procedures, and potential risks (Appendix B). This ensures that participants voluntarily agree to participate, fully understanding the implications of their involvement. The participants always have the right to withdraw from the activities at any time during the experiment, as required by the ethical guidelines provided by the Department of Informatics (2024).

3.6.3 Pre-test Questionnaire

The pre-test questionnaire (Appendix C) acts as a comprehensive data-gathering tool, capturing essential information and contextualizing participants' previous experiences with fraud (Field & Hole, 2003; Oates, 2006). This information aids in understanding the correlation between individual backgrounds and experience with fraudulent content. This part of the procedure is motivated by two aspects: to measure generalizability of the study, and to explore patterns of individual backgrounds on credibility assessment. The former provides the study with insights on whether participant selection has any bias in the findings of the study by examining their individual impact on the external validity, and the latter serves as a way to relate any certain knowledge and experience with affecting an individual's perception of the presented materials.

3.6.4 Randomization

Randomization of exposure is essential to mitigate selection bias and uphold the integrity of the experimental design, as described in Lazar et al. (2017) and Shadish et al. (2003) and discussed in this study’s experimental design section. Through random assignment, each participant is allocated to either the AI-generated or human-made fraudulent content group, ensuring equitable representation across experimental conditions. In this case, the randomization of the content is within-group, which means that no group will solely be exposed to one type of content, or alteration of the independent variable, but instead all of them in a random order.

In practice, the study utilized a constrained randomization technique through software-driven randomization, ensuring each participant received a unique presentation of each material type under varying source and urgency conditions (Lazar et al. 2017).

3.6.5 Exposure to Fraudulent Content

Before exposing the participants with the material, they are briefed with a short description of the content and its context in order to prevent any questions about this and thus a delay that might arise following the initial reaction. In the case of urgency prevalent as per one of the independent variables, participants are encouraged to answer as quickly as possible and to use their intuition, as opposed to the non-urgent setting where participants are motivated to elaborate and take longer to answer. During the exposure phase, the material referenced in Appendix A is presented to the participants, and for each exposure there are several measures (see chapter 3.3.2) taken to capture participants' responses to the fraudulent content. Table 3.2's contents hold the study's dependent variables, along with the corresponding instrument and result metric, and are central to the exposure phase.

Table 3.2: Dependent variables measured with every exposure.

Measure	**Instrument**	**Result metric**
Response time	Stopwatch	Time intervals
Questions and comments	Note taking	Text
Credibility rating	Note taking	Likert scale responses (1-5)

3.6.6 Post-exposure Questionnaire

The post-exposure questionnaire serves as a critical instrument for obtaining participants' reflections and perceptions following exposure to fraudulent content (Lazar et al. 2017; Jarvenpaa et al. 1985). Questions probing experienced levels of difficulty and self-estimated accuracy are provided to the participants and offer a comprehensive understanding of participants' cognitive and affective responses (Appendix D). This information will be used to compare the measured results in order to determine if there is any relation of the participant's own view of their performance and their actual performance. Ultimately, this will serve as an important instrument in assessing if participants have a sound self-understanding of their own capabilities and consciousness of their own perception. Furthermore, is the post-exposure questionnaire concerned with capturing participants' feedback on their experience being a part of this study, allowing the authors to gather valuable insights to improve their future research.

3.6.7 Debriefing Process

At the conclusion of the experiment, participants engage in a debriefing session aimed at clarifying study objectives and outcomes (Field & Hole, 2003; Shadish et al. 2002). Researchers reiterate the importance of digital fraud awareness and provide educational resources to empower participants in safeguarding against future deceptive attempts. This step is also crucial for ensuring ethical safety through its enlightenment that the experiment is solely for scientific purposes, but that the observed phenomena exist in the real-world and can be prevented with knowledge and informative resources. For that reason, participants are in this step handed a guide provided by the European Crime Prevention Network (ECPN, 2022) on how to protect themselves from cybercriminal fraud (Appendix E).

3.7 Data Processing and Analysis

Once the experimental data has been collected, it undergoes processing and analysis to produce meaningful insights and conclusions. This section outlines the steps involved in this phase of the research.

3.7.1 Data Cleaning

Before analysis can be performed, it is important to rectify any errors that may persist in order to minimize the human error factor (Lazar et al. 2017). According to Lazar et al. (2017), this is especially true for any manually entered data, which in this case represents all data collected. Therefore, the authors will conduct a rigorous overview of all entered data to ensure format consistency, accuracy, and appropriate structure.

3.7.2 Descriptive Statistics

Descriptive statistics provide an overview of the basic analysis derived from the collected data, such as median, mean, standard deviation, and frequency distributions. These are calculated to summarize the central tendency, variability, and distribution of the data, and are especially applicable on the more primitive types of data. Together with compelling data visualizations aiding in the interpretation of statistics, these serve as a foundation for further analysis, supporting the identification of patterns and relationships within the dataset (Oates, 2006).

3.7.3 Quantitative Analysis

The quantitative analysis involves using statistical techniques to explore relationships and test previously stated hypotheses (see Chapter 3.2). The experimental design, utilizing within-group repeated measures with two independent and at least three dependent variables, requires the use of a Repeated Measures ANOVA (Analysis of Variance) to assess the effects of different types of fraudulent content on participants' perceptions and responses (Lazar et al. 2017). In cases of missing data or when assumptions of wholeness are violated, more versatile methods like Mixed ANOVA will be employed. This statistical method allows for comparing means across multiple conditions while considering the correlated nature of repeated measures data (Lazar et al. 2017).

This statistical technique helps identify trends based on both dependent and independent variables, normalized among participants (Lazar et al. 2017). If statistical significance is found, Tukey's Honestly Significant Difference (HSD) test will be used for post-hoc comparisons when homogeneity of variance is assumed across groups. When homogeneity of variance is violated, the Games-Howell test will be used instead. Both tests provide pairwise comparisons to identify significant differences between specific groups or measures (Field & Hole, 2003).

3.7.4 Qualitative Analysis

In addition to quantitative analysis, qualitative methods will be used to complement the findings and provide deeper insights into participants' experiences and perceptions measured through unstructured data, such as open-ended responses or categorical data (Lazar et al. 2017). Open-ended responses from exposure activities, as well as pre- and post-questionnaires, are analyzed using thematic analysis to identify key themes. Text analysis, including polarity and subjectivity analysis, and category frequency analysis, are used to understand participants' sentiments and categorization patterns.

However, because the assumptions required for parametric tests like ANOVA are not met with this type of data, the Mann-Whitney U Test will be used instead. This non-parametric test will statistically confirm or refine findings by assessing whether the differences observed between groups are statistically significant (Lazar et al. 2017).

3.7.5 Integration of Results

The results obtained from both quantitative and qualitative analyses are integrated to provide a comprehensive understanding of the research findings. By triangulating multiple sources of data, researchers can corroborate findings, identify converging patterns, and gain a more holistic perspective on the phenomenon under investigation (Oates, 2006).

3.8 Ethical Considerations

The following section will discuss the ethical considerations and implications that this study has on the participants involved in this research, as well as their exposure to content that are designed to mimic cyber threats and frauds. This section is based on the written principles on research ethics provided by the Department of Informatics, Lund University School of Economics and Management (LUSEM) and established by the Research and PhD education committee (FoKom) at the department. In turn, these general principles of what defines good research ethics are mainly based on the Helsinki declaration (World Medical Association, 2013).

3.8.1 Informed Consent of Intentions

Performing an experiment in a laboratory setting, as this study intends to do, involves exposing participants to factors that need to be carefully considered in order to minimize the negative effects on their wellbeing (Oates, 2006; Field & Hole, 2003; Lazar et al. 2017). For this study, the main aspect of consideration is the impact of exposure to potential fraudulent content on human participants. Therefore, it is of the utmost importance to communicate the aim and intentions of the study, and the fact that the participants will be exposed to such material.

Informed consent is obtained from all participants in written form, ensuring that they understand the nature of their involvement and voluntarily agree to participate. Participants are informed of their right to withdraw from the study at any time without penalty. This information will be handed out as a participant information sheet (PIS) based on the content provided by the Department of Informatics' (2017) section on Informed consent and information to research persons.

3.8.2 Confidentiality

All data collected from participants are treated with strict confidentiality. Participants' identities are anonymized, and any personally identifiable information is kept confidential to protect their privacy. Only aggregate data are reported in any publications or presentations resulting from the study, ensuring that individual participants cannot be identified. The above is connected to the participants right to anonymity, as described by Oates (2006) and Field and Hole (2003) as well as required by the Department of Informatics (2024) in order to gain trust and to respect the sensitive nature of private information necessary to collect for the purpose of the study.

3.8.3 Deception

The study involves exposing participants to potentially deceptive content related to cyber threats and frauds. While deception is employed in presenting the fraudulent content, it is essential to minimize any potential harm or distress to participants. Deception is kept to a minimum level necessary to achieve the research objectives, and participants are fully debriefed at the conclusion of the study to explain the nature of the deception and provide educational resources on identifying and avoiding real fraudulent attempts. The above agrees with and is designed to cohere with previously mentioned established ethical guidelines made to eliminate the risk of psychological harm to the participants.

3.9 Validity and Reliability

3.9.1 Generalization and Reproducibility

Human-Computer Interaction (HCI) studies, unlike more deterministic "hard sciences" such as physics, chemistry, and biology, involve measuring human behavior and social interactions which are inherently more variable and hence less replicable (Lazar et al. 2017). Lazar et al. (2017) highlights that these fluctuations, commonly referred to as errors, pose significant challenges in achieving reliable and generalizable results.

The reliability of results in HCI studies is notably influenced by the diversity and number of participants. Achieving representativeness is challenging due to the vast diversity of human nature. Lazar et al. (2017) and Field and Hole (2003) argue that as long as the variability among participant responses is reasonable and the sample size is sufficient to achieve statistical significance, the results can be considered reliable.

Criteria for reliability is defined as the reproducibility and general applicability of the study's findings (Field & Hole, 2003). The study utilizes precise and consistent measurement units such as intervals, Likert scales, and frequency counts, enhancing the reliability of the results. For the more qualitative data that are harder to measure (such as natural text and emotional reactions), established data analysis methods such as sentimental analysis and categorization are used to achieve measurable results. The study's design ensures that as long as the quality of the exposed fraudulent content is consistent and the measurement methods are robust, the findings are reproducible across different samples.

Lazar et al. (2017) separates the role of errors into two types: Random errors and Systematic errors. Random errors are inherent and unavoidable, but their impact should be minimized through careful experimental design and adequate sample sizes. These errors are the ones that need to be explainable by randomness in order to justify the number of participants, else more participants need to be recruited.

To address the systematic errors, the study includes a table (Table 3.3) outlining major sources of systematic errors—such as measurement instruments, experimental procedures, participant selection, experimenter behavior, and the experimental environment—along with the mitigation strategies employed.

Table 3.3: Countermeasures taken to decrease systematic errors (adapted from Lazar et al. 2017).

Systematic error	**Causing factor**	**Mitigation strategy**
Measurement instruments	Appropriacy, accuracy, configuration	Use of extensively tested measurement instruments
Experimental procedures	Planning, participant communication, design	Randomization, established experimental design, written document with participant instructions, pilot study
Participants	Sourcing, representativeness to target group, induced stress during testing	Carefully recruited participants by specified conditions, allow of recover time in between exposures
Experimenter behavior	Scheduling, professionalism, respect	Common sense, consideration of participant wellbeing
Experimental environment	Comfortable furniture, natural settings	Pre-visited and well thought out locations

3.9.2 Internal Validity

Internal validity is concerned with establishing that the observed effects are directly due to the independent variables and no other irrelevant factors (Field & Hole, 2003; Oates, 2006). This study employs randomized within-group experimental designs to reduce the influence of individual participant differences and to ensure that any changes in the dependent variables are directly attributed to the experimental manipulations. This design allows for a reliable way to ensure that the desirable causal relationship is observed (Field & Hole, 2003; Oates, 2006).

According to Field and Hole (2003), measuring a person's behavior in a laboratory setting can itself influence the behavior being measured, thus posing a threat to internal validity. This study addresses such concerns by carefully considering the space in which the experiment takes place in, as stated in Table 3.3.

3.9.3 External Validity

External validity refers to the extent to which the study's findings can be generalized to broader contexts (Field & Hole, 2003; Oates, 2006). The ideal study should reflect general human behaviors and not be confined to specific subsets of the population, or fabricated environments (Field & Hole, 2003; Oates, 2006). This study attempts to achieve external validity by diversifying the participant pool to include students, public sector employees, and private sector employees, thereby broadening the applicability of its findings. To address the effects of the fabricated environment in which the experiment takes place in, independent variables (the fraudulent content) are purely sourced from real-world scenarios, increasing applicability to genuine contexts and not just artificial or non-representative situations. However, it is important to note that laboratory studies inherently cannot replicate real-world conditions perfectly due to the simulated nature of their environments. This limitation is further discussed in below (see chapter 3.10).

3.10 Limitations

The study outlined in this document, while extensive, is subject to several limitations that must be acknowledged. These limitations impact the generalizability of the findings and the potential application of the study's insights to real-world scenarios.

3.10.1 Laboratory Setting

The controlled environment of a laboratory setting allows for the manipulation and measurement of variables in ways that are not feasible in the real world. However, this control also creates a disparity between the experiment's conditions and those of the natural environments where fraudulent content is typically encountered (Oates, 2006; Field & Hole, 2003); Lazar et al. 2017). This limitation is particularly pronounced in the "urgent setting" experiments designed to mimic real-world urgency but still fundamentally different from the unpredictable nature of genuine cyber threats and the stress that might arise naturally in humans.

3.10.2 Material Comparability

One significant limitation arises in the comparison of AI-generated and human-made, specifically in the number of different content sources per material. This variation arises from the unique characteristics of certain materials, such as the absence of human-made equivalents to deepfake fraud. While techniques like image alterations and impersonation, which would be the equivalent to deepfakes, do exist in human-made forms, their quality significantly differs from that of authentic or AI-generated equivalents. Practically, this inequality makes a holistic comparison difficult, leading to an imbalance in the availability of independent variable levels. While such limitations can be circumvented by using statistical tests optimized for missing data points, as well as separating collected data, a holistic and fully comprehensive comparison between AI-generated and human-made content should be carefully interpreted in this experiment.

Lastly, while it is crucial to consider the variation in independent variables to achieve unbiased and balanced results, excluding certain types of cyberattacks due to the absence of human-made equivalents cannot be justified solely on this basis and would distance the experiment from this study’s purpose by delimiting it in a way that could make the results irrelevant or inconclusive.

3.10.3 Replicating the Criminal Mindset

The study's ethical boundaries limit the ability to fully replicate or understand the criminal mindset. Access to real-world cybercriminal tactics, beyond those documented in case studies, is restricted due to ethical considerations. This limitation reduces the depth of psychological and tactical insights that can be incorporated into the experimental design, potentially overlooking strategies employed by sophisticated cybercriminals.

3.10.4 Sample Representation

The willingness of organizations to participate and the variety of participants that can be ethically recruited also pose a challenge, especially seen to the relatively small sample size (n = 15). The study’s sample may not adequately represent the general population or the specific demographics most vulnerable to cyber threats. This limitation affects the external validity and reliability of the findings and suggests caution when projecting the results to broader populations (Oates, 2006; Field & Hole, 2003; Lazar et al. 2017).

4 Results

The purpose of this chapter is to present the results from the experiment conducted to answer the hypotheses stating that AI fraudulent material is perceived as either more or less credible than human-made counterparts, and that the prevalence of urgency either affects attractiveness or argument quality of AI-generated content and therefore influences credibility. The experiment utilizes a within-group design, allowing for randomization of the independent variables, being content source and the prevalence of urgency, and was conducted on 15 participants sourced from the public and private sector, as well as students. The participants were presented with a total of 8 materials, being either human-made, AI-generated or authentic, as well as being either urgent or non-urgent. In turn, credibility ratings and response times were measured along with pre-test and post-test questionnaires collecting information about previous experiences and knowledge, and from a self-assessment before the participants were presented with their results.

The following sections are divided three ways; the first dealing with the first hypothesis (1a; 1b) as well as a detailed comparison of the materials presented, the second presenting the findings of the impact of urgency answering the second hypothesis (2a; 2b), and the third presenting the collected data from the questionnaires and the individual impact of this on participant's performance. Lastly, the main findings are presented, and considerations are brought up to enlighten interpretations of limitations and participant's feedback on the experiment.

4.1 Measured Deception by Content Source

4.1.1 Credibility Ratings

Credibility ratings are measured on a 1-5 Likert scale, where 1 is "Not credible" and 5 is "Credible". Looking at the entirety of the credibility ratings and their reflection on what content source was presented, it is evident that the authentic source had a considerably higher rate (mean = 3.92). As for the human-made material (mean = 2.57), it had a slightly higher rating than the AI-generated ones (mean = 2.29). Interestingly, the human-made materials had a significantly larger spread (std = 1.54) than the AI-generated (std = 1.19), indicating that human-made materials varied more in its perceived credibility, while authentic materials had a lower spread (std = 1.10).

Performing a post-hoc test, in this case the Games-Howell Test, on these content sources by their credibility ratings, in order to achieve pairwise comparisons, it can be concluded that there is no statistical significance between AI-generated sources and human-made ones ($p = 0.730$), with a small effect size adjusted for the small sample size (Hedges' $g = 0.217$). Because the comparison of AI-generated and human-made materials does not show a statistical significance ($p \geq 0.05$), further investigation is needed in order to find considerable differences. Although stating the above results and statistical test, this study fails to reject the null-hypothesis (1a, 1b) and these results should also be concluded as that.

Dividing the materials into two categories, those with three available variations (human-made, AI-generated, and authentic) and those with two (AI-generated and authentic), the results slightly differ but the conclusions remain the same. For materials where there are three counterparts, human-made materials, as expected, remain the same (mean = 2.57), while AI-generated materials increase slightly (mean = 2.46) together with authentic materials (mean = 4.04). Where no human-made counterpart exists, the general credibility surprisingly drops for both the AI-generated (mean = 2.05) and authentic (mean = 3.79) materials. This means that the more complex materials, such as deepfakes and voice impersonations, gained lower trust from participants, and the simpler, but commonly found materials such as phishing emails and websites, and fake news, had a higher credibility. Again, none of these results suggest statistical significance ($p \geq 0.05$) and therefore also fail to reject the null-hypothesis (even with the division accounting for the intentioned bias from the varying levels of the independent variable), and a more detailed presentation of the results is needed to prove AI's capabilities on enhancing cyber fraud on a material specific level.

Further dividing the materials into their respective form, as shown in Figure 4.1, there are crucial observations to be made. It should be noted that when interpreting results at such a detailed level, the reliability of the results decreases due to minimizing the holistic view and should be examined carefully with that in mind. While these results do not directly answer the null hypothesis, as already rejected, they still offer meaningful insights into trends and further investigations. Starting off with the phishing websites, it is evident that AI performs worse (mean = 1.66) on the more complex and detailed websites (*Delivery tracking page*) compared to human-made ones (mean = 3.00), while the simpler websites are identical in performance (mean = 1.6) against the human-made (mean = 1.6) ones (*E-commerce login page*). As for the video deepfakes (*Former US President in video interview* and *CEO talking about a newly launched service*), the AI-generated materials performed considerably worse (mean = 1.86 and mean = 1.71) in deceiving participants than the authentic ones (mean = 4.00 and mean = 4.38), and compared to the general average for AI-generations, suggesting that the complex nature of the video media format is difficult for AI models to replicate. Interestingly, the AI-version (mean = 2.57) of the *Government official delivering a speech to the nation* performed surprisingly well compared to the authentic version (mean = 3.00), indicating that AI excels in voice generation and impersonation. As for the news article material (*News article themed by political tensions*), the human-made (mean = 2.00) and AI-generated (mean = 2.17) versions performed similarly, with the AI-version having a slight lead. The same goes for the emails (*Bank promotional email* and *Streaming service account suspension notice*), where the AI-generated versions (mean = 2.85 and mean = 1.60) are generally judged similarly by credibility, with AI performing slightly better, compared to the human-made (mean = 2.50 and mean = 1.60) ones.

As a last way of categorizing these specific materials, they all utilize different AI functionalities for generating content, either one or more of the following, namely: text, code, video, and audio. Labeling these materials with these functionalities, statistical tests revealed significant differences ($p = 0.000000059$, $p = 0.000046$, $p = 0.000000139$, $p = 0.000004$) in credibility ratings associated with the content source of materials where all AI functionalities had higher credibility ratings for authentic content. While these are rather expected results based on above findings, together with the fact that the human-made and AI-generated materials did not show any significant results when divided by the AI functionalities ($p \geq 0.05$), it is still possible to look at the average ratings in order to derive themes and trends from this data. Comparing human-made and AI-generated materials means of credibility rating categorized by AI functionalities, it is fascinatingly evident that AI-generated materials using code generation is slightly less credible (mean = 2.55) than human-made (mean = 2.71), although in contrast, text-based AI-generated materials are perceived as slightly more credible (mean = 2.15) than the human-made counterparts (mean = 2.07). Comparing this to the control measure being the authentic equivalents, AI-generated versions perform roughly the same in both code and text-generations, with code-generated materials being perceived as 59.4% as credible as authentic ones, and text-based scoring 58.2% on the same measure. Looking at AI functionalities that are not available in human-made versions, the results are less shocking, with AI-generated video content much less credible (mean = 1.79) than authentic equivalents (mean = 4.19) and the same goes for AI-generated audio content (mean = 2.05, mean = 3.79).

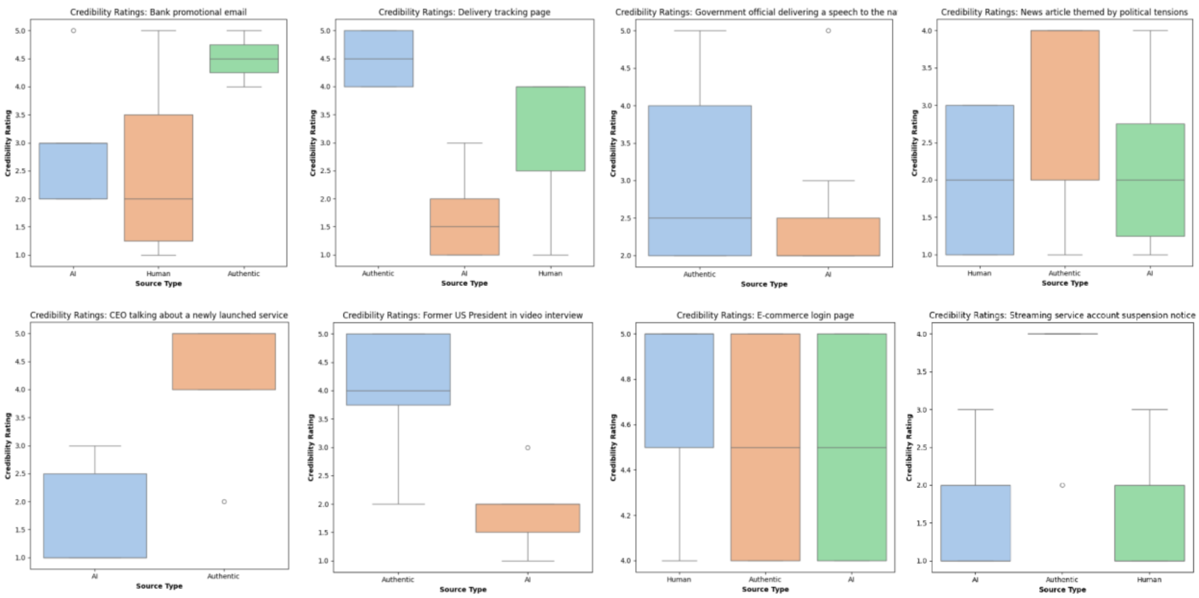


Figure 4.1: Distribution of credibility rating by material and content source.

4.1.2 Response Times

As for the response time, the time it took participants to respond with a credibility rating measured in seconds, the materials performed almost the same, with participants responding slightly faster with AI-generated material presented (mean = 42.04), followed by authentic materials (mean = 45.24) and human-made materials (46.95). Human-made materials had a significantly larger spread (std = 58.61) compared to AI-generated (std = 45.24) and authentic materials (std = 36.68), which could correlate with the same above finding of credibility spread and might indicate that participants found it more difficult to assess the human-made materials compared to the rest. While the response times do not suggest anything besides difficulty or ease of decision, and do not specifically answer any of the hypotheses, they play a major role in assessing the impact of urgency further on.

4.1.3 Comments

The study noted comments that were expressed by participants during the exposure, in order to capture information on subjectivity and polarity for certain content sources and specific materials. Polarity measures the emotional orientation of the text, determining whether the expressed sentiment is positive, negative, or neutral. This metric helps identify the attitude or emotional tone conveyed by the words used in the text. On the other hand, does subjectivity quantify how much of the text is based on opinion, feelings, or beliefs, as opposed to factual information. Assessing subjectivity helps determine how much personal bias or perspective influences the content of the text.

Statistical analyses, the pairwise Mann-Whitney U test more specifically, revealed significant differences in sentiment polarity across different sources, particularly when comparing authentic content with both AI-generated and human-made ($p = 0.0043$ and $p = 0.031$). This indicates that authentic content may significantly influence the sentiment expressed in comments. Examining descriptive statistics, it can be said that, as expected, authentic content on average had more positive comments (mean = 0.006) than the fraudulent versions (mean = −0.0544). In contrast, no significant differences were found in sentiment polarity between human-made content and AI-generated content, although the former had a slightly more positive tone (mean = −0.048) than the latter (mean = −0.061), suggesting that these sources obtained similar responses in terms of sentiment from participants.

Further, the analysis did not reveal significant differences in subjectivity across sources, nor in both polarity and subjectivity across different materials. This suggests that the subjective nature of responses was consistent regardless of the source or material, indicating a uniform perception of content's subjectivity among participants. This uniformity in subjectivity, alongside the noted variations in polarity, provides insight into how different sources may uniquely affect participant perceptions and emotional responses.

It should be noted that these comments were not exact transcriptions of what was said, but instead concentrated notes from participant’s expressions. Therefore, specific word usage or exact citations are unavailable from such data, although a sentiment polarity analysis is still considered appropriate if the general messages of participants are captured properly.

4.2 Impact of Urgency on Credibility

4.2.1 Measure of Urgency

Answering the second hypothesis (2a, 2b), the experiment used the prevalence of urgency to either steer participants into judging the provided material by attractiveness or by argument quality. Some materials inherently possessed urgency cues, such as hurried message of payment, or a pressing call-to-action to change a password and were coupled with instructions to the participant to answer as quickly as possible and to use intuition and first instinct when rating material’s credibility. By adhering to this, a simulated control of the peripheral routes, as stated in the ELM, was made in order to assess the credibility. Ultimately, did this lead to the experiment being able to measure initial attractiveness versus elaborative argument quality for the materials presented.

To begin with, it is evident that urgency played a significant role in dictating the response times of participants, where urgent settings had a faster response time (mean = 26.82) than non-urgent settings (mean = 60.65).

4.2.2 Urgency's Effect by Content Source

Proceeding with answering the question of how AI-generated material performs based on the prevalence of urgency, all three content sources will be presented along with their credibility ratings divided by the either urgent or non-urgent setting, as seen in Figure 4.2. For human-made materials, participants rated them as 25.3% less credible in urgent situations, whereas AI-generated materials surprisingly also were rated as 25.3% less credible in the same urgency setting. For authentic materials, the same metric is 9.4%. This implies two things: that AI-generated material might possess equal initial attractiveness and elaborated argument quality as human-made material, and that they both in general seem to have less content attractiveness than their authentic counterpart. While these results might indicate trends and inspire to draw conclusions for the hypothesis, this interaction effect shows no statistical significance ($p = 0.806$), meaning that urgency doesn't affect credibility ratings differently based on all sources. Although, it can be proven that urgency affects credibility ratings in general ($p = 0.0128$), with urgently presented materials having a lower credibility rating on average (mean = 2.78) than materials presented in a non-urgent setting (mean = 3.25). With these facts laid out, it is evident that the prevalence of urgency cues indeed makes AI-generated content appear less credible, which in turn proves hypothesis 2b.

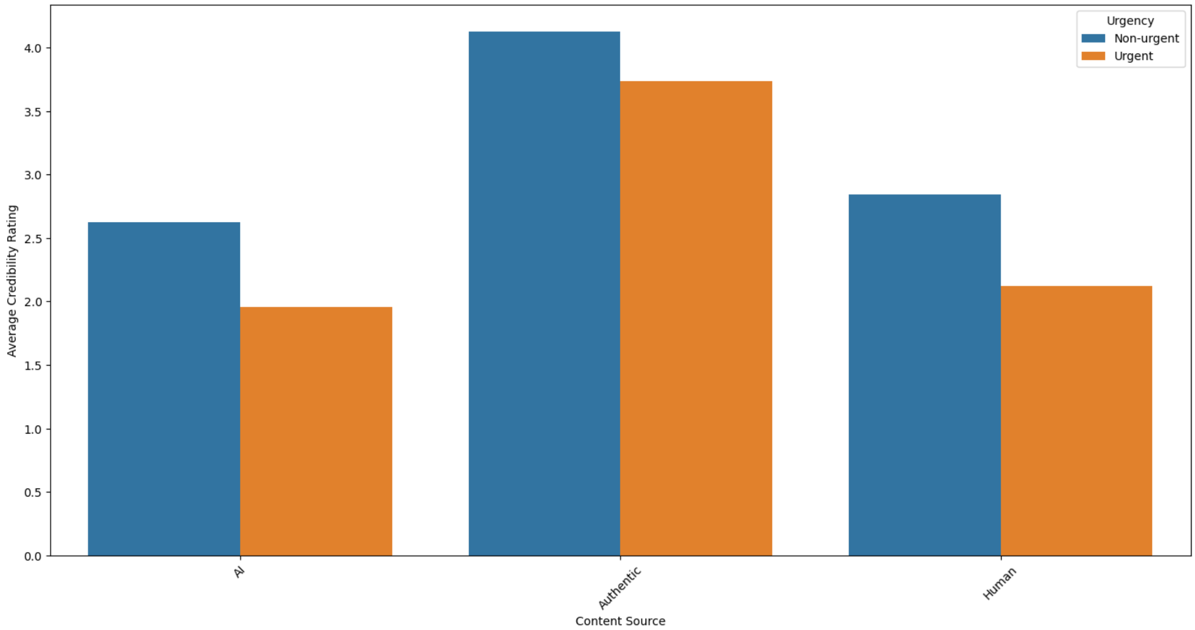


Figure 4.2: Impact of urgency on credibility ratings by content source.

4.3 Self-assessment and Previous Experiences

Before the experiment was conducted, participants answered a pre-test questionnaire with questions regarding their previous experiences with fraudulence and ability to detect fraud, as well as their frequency of use with digital media and their knowledge and experience with AI content. After the experiment, but before participants were presented with their results, they were asked to self-assess their performance, along with their ability to distinguish between AI-generated and human-made fraudulent content, and general feedback about the experiment. Correlations between all quantitative variables collected are presented in Figure 4.3 and further detailed below.

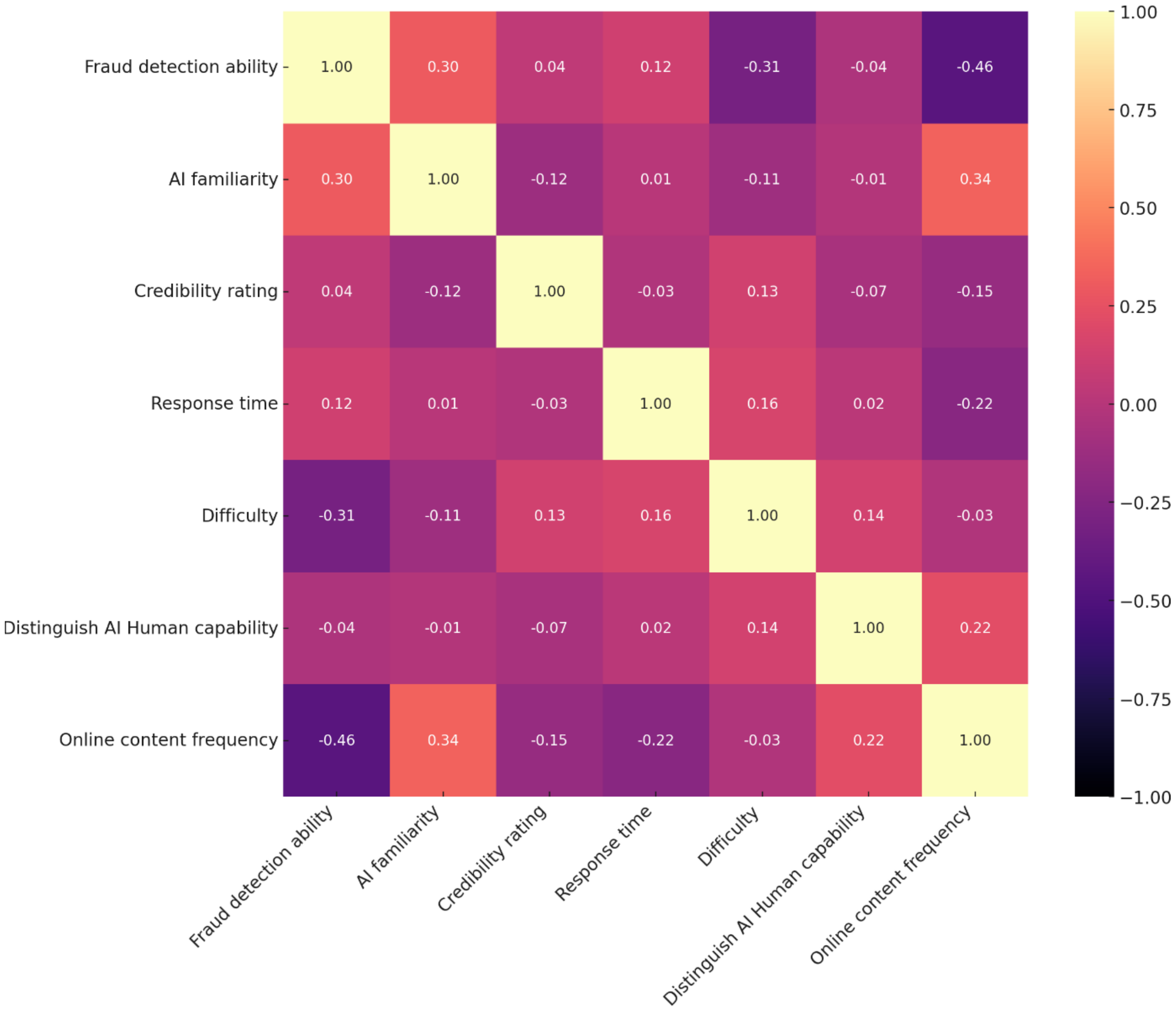


Figure 4.3: Correlation of studied variables.

4.3.1 Correlation of Variables

Participant's fraud detection ability and AI familiarity has a moderate correlation (0.30), suggesting that participants who rate themselves higher in detecting fraud are also more familiar with AI. Furthermore, fraud detection ability and online content frequency have a stronger negative correlation of −0.46, indicating that participants who are more confident in detecting fraud tend to interact with online content less frequently. AI familiarity and online content frequency has a moderate positive correlation of 0.34, suggesting that those who are more familiar with AI also interact with online content more frequently. Credibility rating has very weak negative correlations with both AI familiarity (−0.12) and Online content frequency (−0.15), meaning as familiarity with AI or online content interaction frequency increases, the credibility rating decreases slightly, though these correlations are weak. Response time is not strongly correlated with any other variables, with all correlations being weak. Difficulty of credibility assessment has a moderate negative correlation with Fraud detection ability (−0.31), implying that those with higher self-rated fraud detection ability find the content less difficult to assess. It also has a moderate positive correlation with Response time (0.16), suggesting that as the difficulty of assessing content's credibility increases, response times also increase slightly. Difficulty of credibility assessment and Online content frequency has a very weak positive correlation of 0.14, suggesting a minor increase in perceived difficulty with more frequent interaction with online content. Capability to distinguish AI from human-made content shows very weak positive correlations with Difficulty of credibility assessment (0.14) and Online content frequency (0.22), suggesting that as the difficulty of distinguishing content and the frequency of online content interaction increase, so does the self-assessed capability to distinguish between AI and human content.

4.3.2 Regression Analysis Based on Previous Individual Abilities and Familiarities

In this section, we used regression analysis to explore how pre-test measures (such as fraud detection ability and AI familiarity) predict credibility ratings during the exposure phase of the experiment. The goal was to examine the influence of participants' prior knowledge and experience on their performance in correctly assessing content as either fraudulent or authentic.

The regression analysis reveals that neither fraud detection ability nor AI familiarity significantly predicts the credibility ratings assigned to content. The negative coefficient for AI familiarity (−0.185) suggests a marginal tendency for those more familiar with AI to assign lower credibility ratings, though this trend is not statistically significant ($p = 0.146$). Conversely, the positive coefficient for fraud detection ability (0.185) indicates a slight likelihood that lower ability corresponds with higher credibility ratings, yet this result is also not statistically significant ($p = 0.408$).

4.3.3 Matched Analysis on Content Difficulty Categorization

Looking at the descriptive data of the participant's assessment of 'easier' and 'harder' material, not only a predictive analysis on credibility ratings is suitable, but also a shallower inspection on the participant's self-reflected experiences of the exposure can lead to valuable insights. As for the self-reported harder materials to assess, the distribution is relatively evenly spread, with emails topping the list (6), followed up equally by texts (5), videos (5) and voice recordings (5), and lastly websites (4). As for the 'easier' deemed materials following a less balanced distribution, videos (8) top the list, followed distantly with emails (5), then by voice recordings (4), and lastly by texts (3) and websites (3). These results indicate that while videos for the majority being categorized as an easier material to assess (8), it was also one of the harder to determine for others (5).

Using a paired t-test on a mapping between the exposed materials and the material groupings used in the questionnaire enables comparison of the mean credibility ratings of materials deemed 'easier' versus those considered 'harder' by each participant. The t-test statistic (−1.6583) being negative indicates that the mean credibility rating for 'easier' materials is lower than that for 'harder' materials. The magnitude of the t-test statistic reflects the difference in means relative to the variability in the data. Furthermore, is it evident that there is not a statistically significant difference ($p = 0.1195$) in the credibility ratings for 'easier' versus 'harder' materials based on the data provided. It's important to note that while the test suggests no significant difference at the conventional threshold, it does not prove that there is no difference at all. While the analysis failed to provide statistical proof, there are still themes and trends that can be derived from this data.

4.3.4 Effects of Previous Experience with Fraudulence

Dividing results into two categories, those from participants who have previous experience with fraudulence, by having been a victim of fraud, and those who have not experienced fraudulence by being victimized, conclusions can be drawn on whether this metric affects the perception of online material and if it has any impact on deception. The results show that participants with previous experience with fraudulent content gave a higher credibility rating on average (mean = 3.17) compared to those without such experience (mean = 2.92). Connecting to this, were response times longer for participants with prior fraudulent content experience (mean = 47.73) compared to those without (mean = 41.07).

4.4 Findings

4.4.1 Relation to Hypotheses

Regarding Hypothesis 1a and 1b, which concern the content source and credibility, the empirical investigation did not reject the null hypotheses, demonstrating no statistically significant difference in the credibility assessments between AI-generated and human-created materials. These results suggest that, within the limitations of this study, AI-generated content is perceived with similar credibility as that generated by humans.

In relation to Hypothesis 2a and 2b, which examine the impact of urgency on credibility, the findings indicated that the overall impact of urgency did not significantly differentiate credibility ratings based on if the content were AI-generated or human-made. However, it was observed that materials presented with urgent cues were perceived as less credible across all types of content. This supports hypothesis 2b and implies that urgency influences the credibility perception, likely due to participants employing peripheral route processing under urgent conditions.

4.4.2 Unexpected Findings

The study revealed a notable variability in response times, particularly with human-created materials, which exhibited a wider variability in response times indicating a higher level of difficulty in assessment, potentially due to their heterogeneous nature. Additionally, AI-generated content underperformed in complex scenarios such as deepfake videos, highlighting current technological limitations in creating highly realistic AI outputs in intricate media formats.

Furthermore, did individual differences not show any statistically significant effect on credibility ratings or the prevalence of urgency, and the correlations found between the self-assessments and reported knowledges and experiences were small and non-conclusive. While some correlations were expected, as the one between online content frequency and fraud detection ability, other more meaningful extractions of the results lack significant correlations.

Lastly, the incredibly similar results of the urgency measure between AI-generated and human-made are rather interesting. While the finding itself is not necessarily considered as unexpected, are the closeness of the measures remarkable as it indicates AI’s ability exactly replicate fraudulent attributes very well.

5 Discussion

To encourage the understanding of the discussion that follows, the authors provide a guiding quote that outlines the examined areas and relates them to the argument for their respective findings' importance:

"Identify the current arsenal of cyberattacks that would pose the most harm if AI-generated, assess the capabilities of generative AI that cybercriminals could exploit, and evaluate the possibilities to deceive victims. Compare the usefulness of incorporating generative AI as a tool for cybercrime, and based on this, discuss future actions to mitigate eventual risks."

Table 5.1: Alignment of findings relative to examined study areas.

Area (Chapter)	Finding	Factor
Identification (5.1)	A new dimension of the traditional cyberthreat arsenal	The advanced and generative nature of AI
Capabilities (5.2)	AI's performance in creating deceptive content varies	Multifaceted AI-functionalities
Possibilities (5.3)	AI can mimic the human-like route to persuasion; Individual experiences and knowledge does not affect credibility	Attractiveness and argument quality; Motivation and ability to elaborate
Usefulness (5.4)	AI-generated fraud compares similarly to human-made fraud	Performance in replicating authentic content
Future (5.5)	Efficacy of AI-utilization; Emerge of new techniques; Weak impact of victim's knowledge and experience	Rapid advancement of generative AI; Lack of protective safeguards

5.1 The New Dimension of Cyberthreat Techniques

Identifying the current arsenal of the shapes and forms that cyberattacks come in, especially with the rise of generative AI, is crucial in ensuring that the correct safeguards are applied. This study tested the deception of these innovative approaches such as deepfakes and voice impersonation, and it is already evident that this new generation of attacks exist and that it is posing a threat to traditional prevention systems (Homeland Security, 2022; CISA, 2023; Google, 2024). While this study also tested conventional cyberattacks, such as phishing emails and websites and fake news, it enables the comparison between the both and is therefore able to draw conclusions on what the addition of these new approaches mean for the entire arsenal of cyberattack techniques. Therefore, the following section will divide the examined cyberattacks into two categories: those that are previously unseen and fully made possible with generative AI, and those that existed before that and are traditionally made by humans.

5.1.1 Innovation of New Cyberattacks

Calling these attacks fueled by generative AI would be an understatement, as the authors of this study found it practically impossible for a layman to craft human-equivalent deepfakes and voice impersonations. While deception is achievable without advanced digital skills, this lack of understanding of the real-world and the criminal mindset is indeed a limitation to this study. Although this would nonetheless mean that the power to craft advanced deceptive materials, impersonating influential individuals and institutions, now suddenly lies in the hands of anyone with an internet connection for a small fee. These benefits might be general for all uses with generative AI and will be further discussed in-depth below (see chapter 5.4) and should not be noted as a finding of this study, but rather an informal observation of the research process.

With this as background, this study found that while deepfakes and voice impersonations are non-equivalent to anything human-made with the same skills required, they do not necessarily pose as such a serious threat as conventional, human-made attacks. In this study's context, deepfakes were gathered as spoofing attacks, while voice impersonations were generated to replicate disinformation operations. Participants could rather easily distinguish between the authentic and the fraudulent materials, especially for the deepfakes. While voice impersonations performed surprisingly well at deceiving participants, they were still considered as less credible. This could be due to the personal and complex nature of such materials, being in rich formats such as audio and video as well as impersonating other human individuals. While this study's results could be considered as a snapshot of the present stage in the evolution of deepfakes and voice impersonations, it is important to note the staggering increase in sophistication of these techniques, as noted by the collaborative works of NSA, FBI, and CISA (2023) and feared by tech giant Google (2024) and simultaneously predicted to pose the most harm as per Caldwell et al. (2020). This means that the current technologies to craft completely authentic-looking and sounding frauds might not be there yet, but there is proof that these advancements need proactive safety measures taken to prevent future harm.

5.1.2 Enhancing Traditional Cyberattacks

Purely looking at traditional cyberattacks that exist no matter the existence of generative AI, it can be concluded that these perform better at deceiving participants. It is also clear, though rather a trend than statistical evidence, that simpler, less complex materials that are AI-generated perform generally equally to their human-made counterparts. Taking websites as an example, the results showed an apparent differentiation of participant's credibility ratings on a simple, rather stripped-down log-in page compared to a more complex and detailed parcel tracking page. While these results are inconclusive, they are still themed towards, together with the predictive exponential increase of AI advancements as per Gartner (2024) and UK's National Cyber Security Center (2024), AI potentially being able to handle generating and enhancing complex cyberattack materials in the future.

5.2 AI-functionalities in Supporting Different Cyberattacks

The characteristics and appearance of cyberattacks differ greatly, with some being as simple as an email, and some as complex as a video. This study examined three major attacks: phishing, disinformation operations and spoofing, chosen by their suitability to be enhanced by AI together with their highly predicted harm (Google, 2024; Caldwell et al. 2020). These attacks were then further categorized into 8 materials, namely emails, websites, news articles, voice recordings and video. Each one of these materials is supported by one or more of the selected capabilities of AI, as described by Schmitt and Flechais (2023) and Google (2024) as text, programming code, voice and video. With these categorizations made, this study is able to draw specific connections between the multi-modal functionalities of generative AI with specific cyberattacks.

While the only significant causal relationship observed was between authentic and fraudulent (incorporating both AI-generated and human-made materials), which will be further discussed below (see chapter 5.4), there are still some themes that can be observed. Based on the assumption that text and code are seen as less complex formats compared to audio and video, it is evident that AI-generated materials perform considerably better in simpler formats, and especially when utilizing the text-generating AI functionality. This adheres with the above findings regarding deepfakes and voice impersonations. Ultimately, this means that generative AI may generally fail at enhancing complex cyberattacks, such as spoofing attacks and disinformation operations incorporating deepfakes and voice impersonations but performs slightly better assisting in crafting simpler and human-comparable content for cyberattacks, such as phishing emails. This aligns with the findings of Zhai et al. (2023) where text-generated content performed better than that of human-made. While the difference found is marginal, and without statistical significance, it is still a trend that can be confirmed by existing research.

Furthermore, this finding aligns with existing research stating that a primary benefit of using generative AI to craft cyberattack content would be by mitigating the human error (Google, 2024; Mitchell, 2019), which is apparent when this study concludes that examined AI-generated materials existing in human-made equivalents performs better than those more complex ones only existing in AI-generated versions. While this finding simultaneously contradicts previous research's predictions of generative AI's power in crafting deepfakes and other complex media, it should be noted that these are forecasts for the future that might not be true just yet (Google, 2024; Caldwell et al. 2020).

5.3 Persuasive Routes and Victim's Perception

5.3.1 Replicating the Human-like Route to Persuasion

A substantial part of this study is related to examining the way AI-generated fraudulent content is perceived by victims, in order to understand its effectiveness and thereby its strengths and weaknesses. This is done in order to lay the fundamental groundwork for informing policy makers on how the preventive measures must be designed in order to mitigate any potential threat that generative AI might possess in assisting in cybercrime. To do so, this study has applied the Elaboration Likelihood Model (ELM), as proposed by Petty and Cacioppo (1986), to supportively inform this study of the cues and routes that affect the victim's decision to trust or disbelieve specific content, such as emails, websites, videos, and audio. In practice, this study exposed participants of the experiments with two types of each material: an urgent and a non-urgent version. The urgent version emphasized the peripheral route and relied on attractive attributes to alter the participant's perception, relying on their intuition and instinctive guidelines by motivating participants to answer as quickly as possible and as soon as they had an initial answer. The non-urgent counterpart encouraged participants to carefully examine the material and reason its authenticity with logic, by elaborating on its argumentative quality.

The study achieved this distinction between the versions, seeing to the significant differences in response times and credibility ratings of the urgent and non-urgent versions. But most interestingly, did the AI-generated fraudulent materials take the exact same route to persuasion as the human-made fraudulent ones - and not like the authentic counterparts. The results do not even differ by a tenth of a percent, as opposed to the authentic versions having a significantly different route of persuasion. This could imply that the generative AI models used to craft the fraudulent materials did not try to replicate the authentic versions, but instead actually, and effectively, aimed to mimic fraudulent material. This phenomenon could be confirmed by examining the insides of how AI works, and that it consists of weighted algorithms trained on massive amounts of data as described in the works of Hubert et al. (2024). With this understanding in place, the result of this study implies that generative AI could be trained on fraudulent material, and that the AI models might have known the intentions of the instructions, without them being explicitly told, and provided output that is derived from the fraudulent material it was trained on. While this speculation is loose and far-fetched, what is surprising about this is that none of these prompts (Appendix A) instruct the AI model to replicate anything fraudulent or counterfeit, but instead prompted it to generate authentic material. Knowing that generative AI takes prompts as its input, allowing the user to adjust the output as preferred, as the authors have done in this study in order to generate the materials exposed in the experiment makes it a rather strange observation (Hubert et al. 2024).

Ultimately, this study found that with the prevalence of urgency, all of the materials, no matter if they were AI-generated, human-made or authentic, decreased in credibility ratings. This means that all materials presented had a better performance when elaborated on using a systematic approach rather than using cognitive heuristics with limited ability and motivation to apply reasoning logic, as per the ELM, Stalans (2021) and Metzger and Flanagin (2013). What differs between the authentic and fraudulent content, is that fraudulent content has higher argumentative quality, decreasing in credibility even more than authentic materials in urgent settings. This finding offers valuable insights in how AI-generated frauds are processed and evaluated by victims, allowing for a greater understanding of its implications in the cybersecurity landscape. In practice, this would mean that policymakers and AI service providers can focus on mitigating AI's strengths in supporting in the craftsmanship of fraud by knowing how it affects perception, and therefore potentially effectivizing their effort in building safer security measures that restrict unethical use and ultimately protect victims of enhanced harm.

It should be said that the urgent independent variable was pre-defined for every type of material, meaning that the prevalence of urgence was a fixed setting and did not change for the same materials. This means that these results could be affected by the type of materials that were urgent and non-urgent, and not the sense of urgency itself. Although this limitation is consequential by allowing for bias in the results, it is hard to believe that it can account for such a considerable finding.

5.3.2 The Minimal Impact of Individual Background

Before the experiment, participants were made to answer a questionnaire, asking questions about their previous experiences with AI, fraud and online content, as well as their self-rated ability to detect fraud. After the experiment, participants were asked to rate the difficulty of assessing the credibility of the materials presented, along with a question about their ability, if any, to distinguish AI-content from human-made content. This information is valuable in understanding the implications of individual factors on victim's perception, as perception not necessarily is something general for all humans but instead can be argued to build upon previous life-experiences, skills, and knowledge (Metzger and Flanagin, 2013; Stalans, 2021).

The correlations indicate that pre-exposure factors do not strongly predict how participants rate the credibility of content or their response times. The same applies to post-exposure reflections, where difficulty and self-assessed capability to distinguish AI-generated content from human-made content show only weak correlations with actual performance metrics. The strongest relationship observed here is between online content frequency and response time, suggesting that those who interact with online content more frequently may respond quicker to the tasks in the experiment. Another expected finding is that participants who reported a lower ability to detect fraud generally found the experiment to be more difficult.

Overall, the regression analyses suggest that the pre-exposure factors of fraud detection ability and AI familiarity do not significantly predict how participants rate the credibility of content. This result most notably makes the above correlational findings irrelevant and points them to being interpreted as more of a trend rather than hard evidence. Moreover, does these results indicate that other factors, potentially those related to the situational context or specific characteristics of the content, play a larger role in shaping these outcomes during the experiment. These insights underscore that mere awareness or familiarity with AI does not necessarily equate to an enhanced capacity to critically evaluate or swiftly respond to AI-generated content, highlighting the complexities inherent in human-AI interactions. While not statistically significant, some of the findings of the regression analyses do point to certain trends, such as higher AI familiarity having a slight effect on lower credibility ratings, signifying that awareness of the capabilities of generative AI potentially makes individuals more suspicious of potential fraud. Another trend also found in the analysis might be that lower fraud detection ability among individuals corresponds with higher credibility ratings, possibly indicating that the lack of this ability makes victims more trustful to potential fraud, although this correlation is marginal.

Lastly, an observation could be made regarding participants who have been victimized by fraud before. Those who answered positively showed higher average ratings of credibility, as well as took longer to respond. This could either mean that victims who have been victimized before are more likely to be victimized again, or that their experience in what constitutes fraud leads to more informed decisions on correctly identifying content as either authentic or fraudulent. This finding relates to Stalans' (2021), suggesting that prior victimization is a complicated matter often affected by emotional connections to the incident and misinterpretations of the underlying contributions to the susceptibility of the victim.

Ultimately, this study found little to no evidence of the relation between individual experiences and fraud, as well as self-assessed metrics on performance and actual performance. This mostly contradicts the works of Metzger and Flanagin (2013) by proving that none of the varied metrics on individuality had any significant impact on cognitive heuristics. However, it is important to note that Metzger and Flanagin (2013) have no outspoken model on how certain aspects of background affect perception, but instead argue for the fact that knowledge and experiences affect credibility in some way or another. While this finding is positive for the study’s generalizability for a broader population as it serves as important evidence for the nonexistence of bias in the participant selection, it is again important to note the small sample size (n = 15) and therefore the high risk of random errors. However, if these findings are proven to be true, this has serious impacts on the above-mentioned safety measures that need to be taken to protect victims, as it is then evident that victims have little to no effect on taking proactive measures to protect themselves, putting a higher strain on those responsible for ensuring ethical use of generative AI services.

5.4 Effectiveness of Utilizing Generative AI in Crafting Cyberattacks

The following discussion will emphasize the implications that generative AI has in supporting cybercriminals in crafting fraudulent content, and besides the main results of credibility it will also focus on the broader aspects that affect the usability of AI for this purpose.

As expected, this study found a significant difference between the perceived credibility of authentic versus fraudulent materials. Simultaneously, it found no evidence that credibility differs when AI-generated or human-made materials are assessed. While this might sound like inconclusive results, it actually means that AI-generated frauds perform roughly the same as those generated by humans. Furthermore, it is even evident as discussed previously that these materials are not only similar in credibility ratings, but even further alike in the way that they are perceived by participants. In practice, this would mean that cybercriminals have little to no effect crafting the content themselves compared to using generative AI to craft fraudulent content. As discussed previously, certain AI functionalities point to a trend in creating content that exceeds that of human capacity or human-made quality, such as deepfakes, voice impersonations and simpler, less complex content, but central to this present discussion is the actual usability of generative AI – which the study found to be equal than that of human-made equivalents.

However, the context of using generative AI for malicious purposes needs to be discussed in order to determine the overall usability. While this study found no direct relation to generative AI being able to produce content better than that of cybercriminals themselves, the availability, scalability and efficacy that is inherent in generative AI are of great importance when determining the usability at large. Caldwell et al. (2020) express concerns regarding the potential for automation and scalability when using generative AI for cybercriminal purposes, as the digital nature of this technology could allow for faster times to generate. Furthermore, Schmitt and Flechais (2023) are particularly concerned about the potential in social engineering that generative AI possesses because of its ability to be trained on data, which in turn could allow it to pose a serious threat when learning about a victim’s personal information used in social engineering attacks. As for the availability of AI services that can support unethical activities, there is no doubt that these are available to anyone for use, as observed by Falade (2023) and Gupta et al. (2023). To these findings, the authors of this study are willing to agree, although an informal finding to the results, they perceive AI services to be surprisingly easy to use for crafting unauthentic content. As a last note, none of the above findings are directly central to this study, covering extended capabilities of AI further than that of generating fraudulent content, but are crucial to have in mind when interpreting this study’s findings.

Compiling existing research, it appears that most studies found AI-generated content to perform moderately overall. Most of the research comparing human-made and AI-generated output focuses on text-based content, while this study focused on the broader capabilities as well. This study found no significant differences between AI-generated and human-made content in terms of credibility for fraudulent purposes. This finding directly contradicts Zhai et al. (2024), who suggested otherwise, and indicates that the expectations of generative AI, as proposed by Morris (2024), might not yet be realized. Furthermore, these results partially diverge from those of Májovský et al. (2023), who concluded that while AI-generated content initially performs well, it lacks the factual accuracy and logical coherence required for authenticity upon closer examination by the viewer. Although Májovský et al. (2023) acknowledged the strengths and weaknesses of AI-generated content as stated, they also noted that in general, AI performed comparably to humans, thereby linking their findings with those of this study. It is important to note that this study was conducted among experts, in contrast to the broader participant selection in our study. Lastly, there are intriguing links between these findings, such as AI's slight advantage in the ability to generate simpler media like text and the observation that AI-generated fraud performs roughly on par with human-generated fraud in general.

5.5 The Need for Mitigating Harmful Trends

This section of the discussion will go over the implications of this study's findings on cybersecurity, and how these findings can be applied to improve safety measures in generative AI services by suggesting how policies such as those made by UNESCO (2024), OECD (2019), and ISO (n.d.) can evolve to better govern the use of AI in crafting fraudulent content as suggested by Ferrara (2024). Besides from laying the fundamentals of policies and regulations, the findings will also be discussed in the perspective of the greater society, in order to also inform individuals and businesses on the current state of AI-driven cybercrime.

This study found no considerable differences in the perceived credibility between human-made and AI-generated frauds. In practice, this implies that cybercriminals do not have any advantage nor disadvantage seen to the success-rate of their attacks by using generative AI. However, the contextual advantages such as availability, scalability and automation of using AI in crafting fraudulent content is still considered as a concern (Caldwell et al. 2020; Falade, 2023; Gupta et al. 2023). The rise of unprecedented techniques like deepfake media and voice impersonation presents new threats. These techniques could be exploited for disinformation campaigns and spoofing attacks. Furthermore, generative AI has the potential to assist cybercriminals in certain scenarios, given its rapid development. It's clear that policies and internal safeguards must adapt to these accelerating advancements to proactively address these harmful trends. Moreover, this study found little to no evidence of individual's experiences and knowledge affecting their performance in correctly identifying fraud, indicating that an even greater responsibility is put on AI policymakers and service providers to protect potential victims from fraud. The findings on prior victimization adheres with Stalans' (2023) suggesting that learning from a fraudulent experience may be influenced by emotions and misunderstanding, but simultaneously contradicts that of the broader psychology scope as stated in the ELM and by Metzger and Flanagin (2013) pointing out that individual differences in terms of knowledge and habit affects perception. The fact that the results are mixed are of no surprise connecting them to Stalans (2023) discussion on that existing research is divided in this topic.

To draw practical conclusions from how these findings can be used to better steer focus onto concerning areas in the malicious use of generative AI, this study will hopefully serve as a fundamental provider of information in order to make decisions on how to improve existing policies and regulations. The findings of this study, particularly in the context of existing research, suggest that the existence of the policies made by UNESCO (2024), OECD (2019), and ISO (n.d.) are justified, but even more so that they need to keep getting evolved to tackle this ever-changing AI technology, as pointed out by Ferrara (2024). While these exact measures are out of the scope of this study, it will hopefully shed some light on the implications that AI-generated fraud have on victim perception for further research to determine the precise actions needed to prevent this harmful trend.

5.6 Considerations on Interpretation

The following section will discuss the impact of the study's identified limitations (see chapter 3.10) on the observed results, and how these limitations affect the interpretation and drawing of conclusions based on the findings.

5.6.1 Laboratory Setting: Impact on Credibility and Urgency

The experimental design conducted in a laboratory setting presents the primary limitation of this study stemming from the methodology applied (Lazar et al. 2017; Oates, 2006; Field & Hole, 2003). Examining the control variable used in the exposure, the authentic content, it is evident that participants were affected by artificial setting and the suspicion awareness of fraudulent content potentially present. Participants sometimes also found the authentic material less credible than expected. While this effect is accounted for in the experimental design, it is a notable observation that could potentially affect other aspects of the results. One such aspect that needs to be addressed is the simulation of the urgent setting conducted in the experiment. The urgency variables were meant to steer participants into using their intuition and therefore relying on heuristics, as this is a tactic often used by cybercriminals (Stalans, 2023). The response time variable was used as a metric to guide participants into using less elaboration and therefore simulating cognitive heuristics, as per Petty and Cacioppo (1986) and Stalans (2021) in the ELM, and this study found it evident that seen to this metric the urgent simulation was successful. Although while the difference in response time observed between the urgency settings in this experiment could be argued to simulate this phenomenon, it is still unknown to which extent this setting affected participants' actual emotions and use of perception on a psychological level, as there are multiple factors that decide what cognitive style individuals take more than just by time pressure (Stalans, 2023).

5.6.2 Material Disparity: Impact on Statistical Bias and Holistic Results

Another critical limitation involves the comparability of AI-generated and human-made fraudulent content in certain materials. This study generally found materials not having any human-made counterparts, due to the nonexistence of qualitative comparisons, were generally rates as less credible than those with all three variations of the material. While this lack of data was avoided using statistical methods as per Lazar et al. (2017) and a detailed presentation of the results, a holistic view of the general results is still skewed by this limitation. Arguing for including these lesser quality materials in the study could have affected the results by lowering the general credibility of human-made content, which is an important consideration to have in mind when interpreting the results.

5.6.3 Insufficient Demographic Data: Impact on Representation

The sample size and diversity of participants also impose significant constraints on the study. While the results did not show any significant differences between participants self-reported individualities and their performance in assessing credibility, it should be noted that limited information were collected about participants demographic background, which could have potentially yielded different results. Looking alone at the individual information collected from participants, this study is arguably able to represent the broader population, but the lack of results stemming from a richer analysis on participants' background should be noted.

5.6.4 Feedback from Participants

The feedback expressed by participants was solely positive, being themed towards appreciation of the study's educational impact and its engaging subject matter. Many expressed gratitude, such as one participant who mentioned, "Thanks for increasing my awareness of this potential threat", highlighting the educational value of the study. Others found the subject matter captivating, with remarks like "Very interesting subject", showing the relevance and interest generated by the topic. These responses indicate the participants' recognition of the study's significance and their improved understanding of cybersecurity threats posed by AI-generated fraud.

6 Conclusion

This thesis aimed to investigate the impact that generative AI has on cybercrime, particularly in the creation of fraudulent content. Drawing on a variety of previously existing scientific studies, along with this study's own experimental data, a comprehensive understanding was provided of how AI-generated fraudulent content compares to its human-made counterpart in influencing victims' perceptions of credibility and susceptibility to fraud.

Addressing the first research question, the research findings suggest that AI-generated fraudulent content does not significantly differ in perceived credibility from human-made content. Both types of content have similar impacts on victims, challenging the idea that AI inherently enhances the deceptive quality of fraudulent material. This finding raises important questions about the current and future capabilities of AI in cybercrime, suggesting that while AI may not yet surpass human capabilities in crafting deception, it can equal them in the efficiency of deception. Interestingly, it is evident that fraudulent content generated by AI takes the same route to persuasion as that of human-made equivalents, as opposed to the significantly different way that authentic content affects victims' perception. More specifically, these findings suggest that AI-generated content, and thus human-made, are perceived as having a wider gap between higher argumentative quality and lower initial attraction, compared to authentic content, which has a much narrower gap (Figure 4.2). Further fascinatingly is that this study also concluded that participants' self-assessed abilities, along with their previous experiences and knowledge, do not impact their perceptive performance in correctly assessing content as being either authentic or fraudulent. This indicates that policymakers and AI service providers have a higher responsibility to ensure that AI is used for ethical purposes, restricting its harm on individuals.

Answering the second research question, this study addressed how AI enhances cybercriminals' capabilities in creating fraud, considering the rise of new forms of fraud and the accelerating technological innovation in AI. The use of generative AI in cybercrime introduces a range of previously unseen methods for executing cyberattacks, such as deepfake technologies and voice impersonations. However, the effectiveness of these AI-driven strategies, while potent in theory, has not shown a distinct advantage over traditional methods in practice, except in their ability to be inequivalent to anything seen before crafted with the same skills and engagement by humans. Moreover, existing research is particularly concerned about other aspects inherent in AI as a technology, such as availability, scalability, and automation, that might not directly affect the attack success rate but instead offers a whole new world of possibilities for facilitating criminal activities (Caldwell et al. 2020; Falade, 2023; Gupta et al. 2023). Due to this, it could be argued that cybercriminals still might benefit from using generative AI, as these contextual aspects might help outweigh traditional methods.

Though not significant nor reliable, certain trends can be observed through these findings, such as AI-generated simpler and less-complex content performing slightly better than human-made equivalents, such as text. For other materials, such as detailed websites, it can, on the contrary, be observed that human-made materials perform better, suggesting that generative AI lacks the ability to handle richer media formats.

Our analysis, supported by a thorough literature review and experimental evaluations, indicates that the current state of AI-driven cybercrime involves a complex interaction of technological capabilities and human vulnerabilities. While AI offers a new dimension of fraudulent media with the introduction of deepfakes and voice impersonations and can potentially optimize certain aspects and types of more conventional cyber fraud such as phishing, the fundamental dynamics of deception and victim response remain comparatively the same as traditional methods. Though these might sound like inconclusive findings, they instead practically imply that cybercriminals no longer have any reason to craft fraudulent materials themselves anymore, as AI-generated works are equally as effective. In light of existing research, predictions indicate that generative AI will pose an even more serious threat to cybersecurity in the future (Ferrara, 2024; Google, 2024; Caldwell et al. 2020). Together with this study's findings, it is clear how and to what extent generative AI affects victims' perceptions of credibility and their susceptibility to fraud. This calls for proactive measures in the cybersecurity field, by pushing on regulators, policymakers and AI service providers to ensure safe AI systems and to individuals and organizations to stay informed on the current state of this concerning trend.

6.1 Further Research

The findings of this study call for several areas for further research to expand our understanding of generative AI's role in cybercrime. An important area to explore is the broader use of AI in social engineering tactics. Future studies could investigate how generative AI may influence more advanced aspects of fraud, such as the learning of the victim's personal lives to personalize attacks. Understanding the psychological impact of these AI-generated interactions on human decision-making is also crucial. Furthermore, would it be interesting to examine the long-term impact on trust in the general population stemming from the introduction of generative AI in cybercrime, as this technology arguably complicates the detection of online fraud and could potentially damage relations with genuine communications.

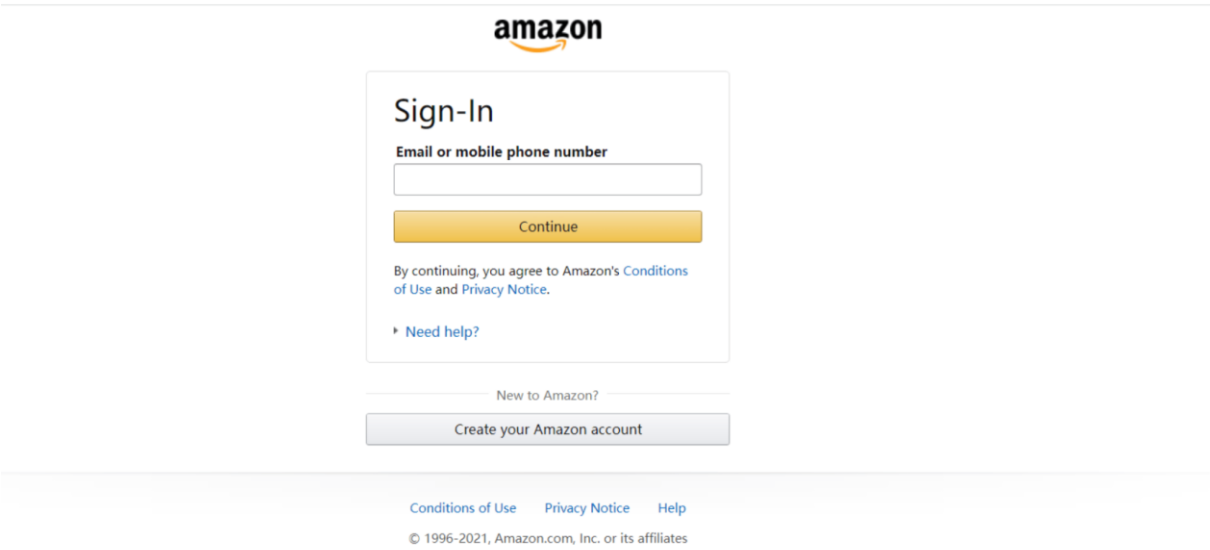
Additionally, conducting larger-scale studies that include varied populations and more materials could help verify the generalizability of the findings and uncover how different cultural and demographic groups react to different AI-generated content cyberattacks. This research could extend to developing and testing new cybersecurity measures and educational programs specifically designed to combat the unique challenges posed by generative AI in cybercrime.

Appendix A - Materials

Phishing website content

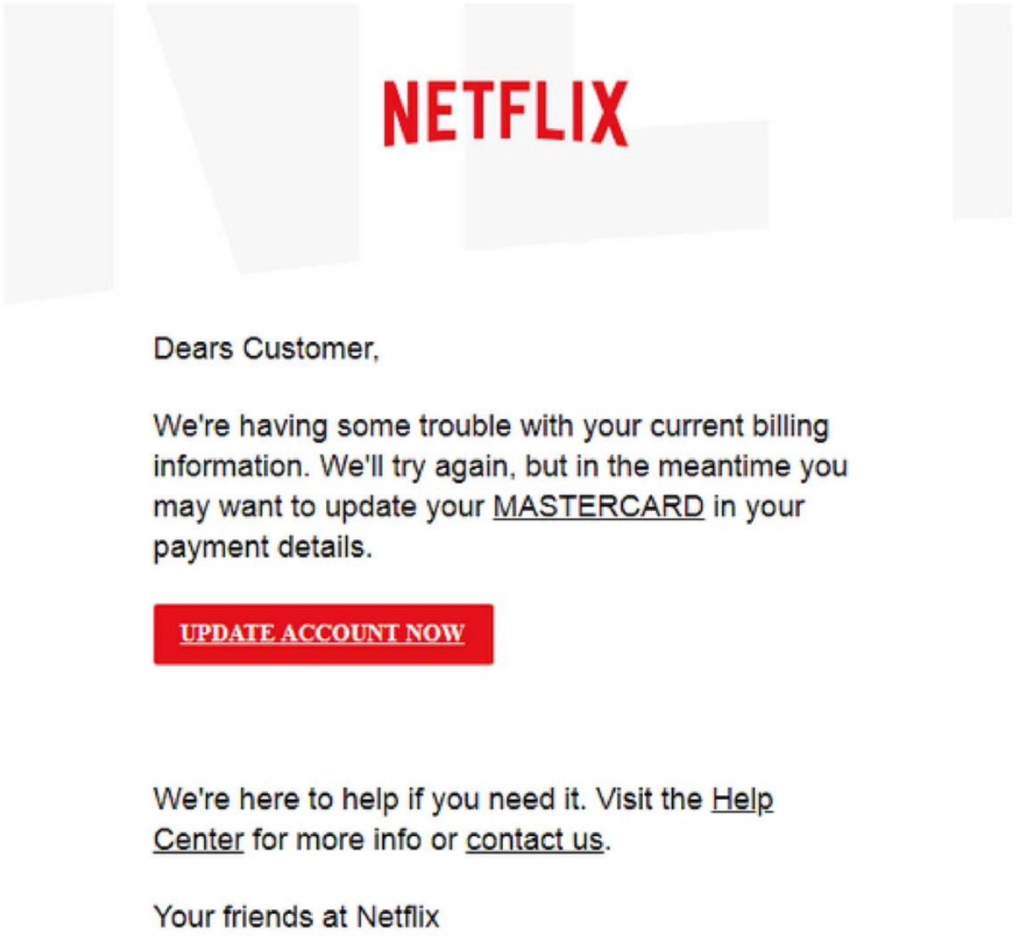
Human-made Fraudulent Version – E-commerce Login Page

Source: *Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning* (https://www.mdpi.com/sensors/sensors-21-08281/article_deploy/html/images/sensors-21-08281-g001.png)



Human-made Fraudulent Version – Streaming Service Account Suspension Email

Source: Mailguard (https://www.mailguard.com.au/blog/netflix-spoofed-once-again-in-phishing-email-scam)



NETFLIX

Dears Customer,

We're having some trouble with your current billing information. We'll try again, but in the meantime you may want to update your MASTERCARD in your payment details.

UPDATE ACCOUNT NOW

We're here to help if you need it. Visit the Help Center for more info or contact us.

Your friends at Netflix

Human-made Fraudulent Version - Bank Promotional Email

Source: Säkerhetskollen (https://sakerhetskollen.se/aktuella-brott/phishingmail-som-utnyttjar-nordeas-namn)

NORDEA

Med det nya **Nordea BankID-** systemet kan du använda alla våra nätbanksfunktioner på din mobiltelefon. Hantera dina konton, kreditkort, insättningar och ekonomi var som helst. Upplev den fullständiga direktbankstjänsten med de flesta gratis konton i Sverige på din enhet.

FULL TILLGÅNG till konton, kort, finansiering och depå samt alla webbbanksfunktioner.

Den **Nordea BankID** erbjuder ett komplett utbud av bank- eller tjänster Nordea. Med den kan du enkelt hantera ditt kostnadsfria löpande konto, kreditkort, samtalskonton, finansiering och dela depå. Särskilt praktiskt: Designa din ekonomiska status individuellt – så optimerar du åtkomsten till dina konton.

.

Aktivera nu " Nordea BankID" genom att följa instruktionerna:

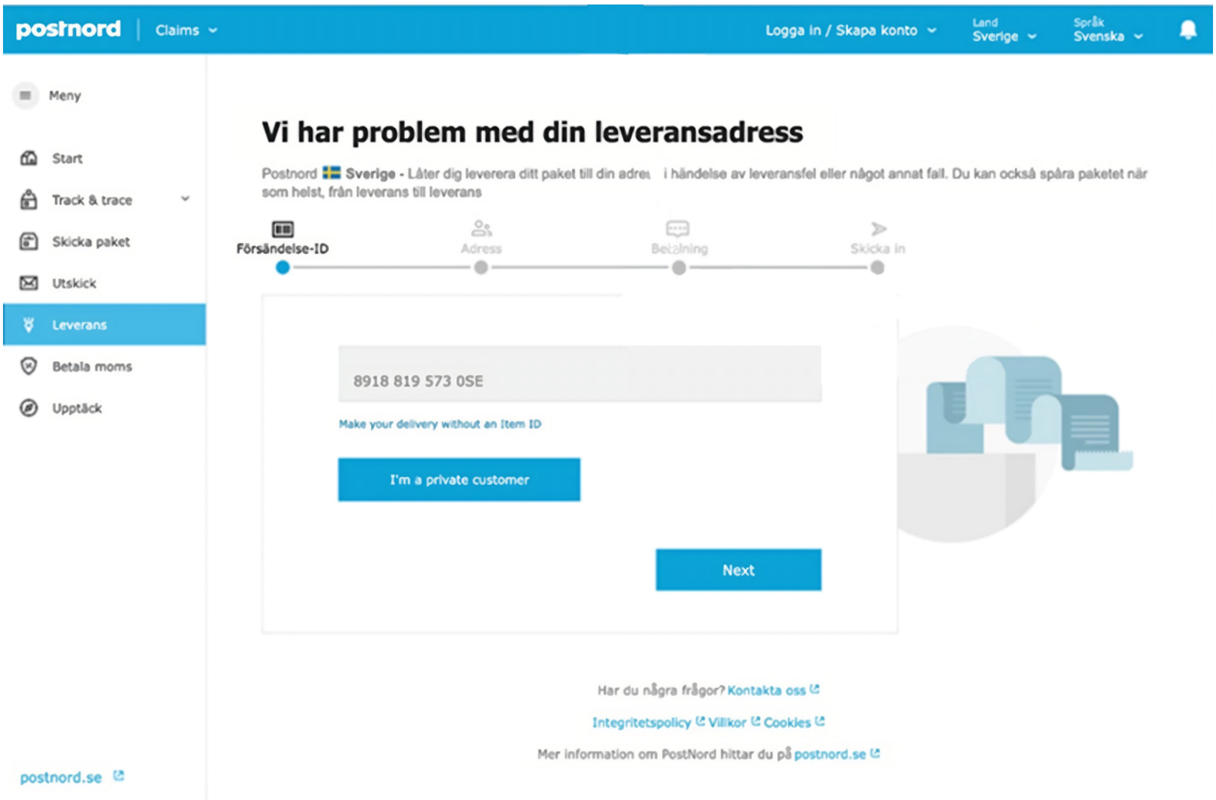
1) Identifiera dig med dina bankuppgifter.

2) Bekräftelsen av din **e-Kod** har genomförts framgångsrikt.

Med vänliga hälsningar,
ditt Nordeas service team.

Human-made Fraudulent Version - Delivery Tracking Page

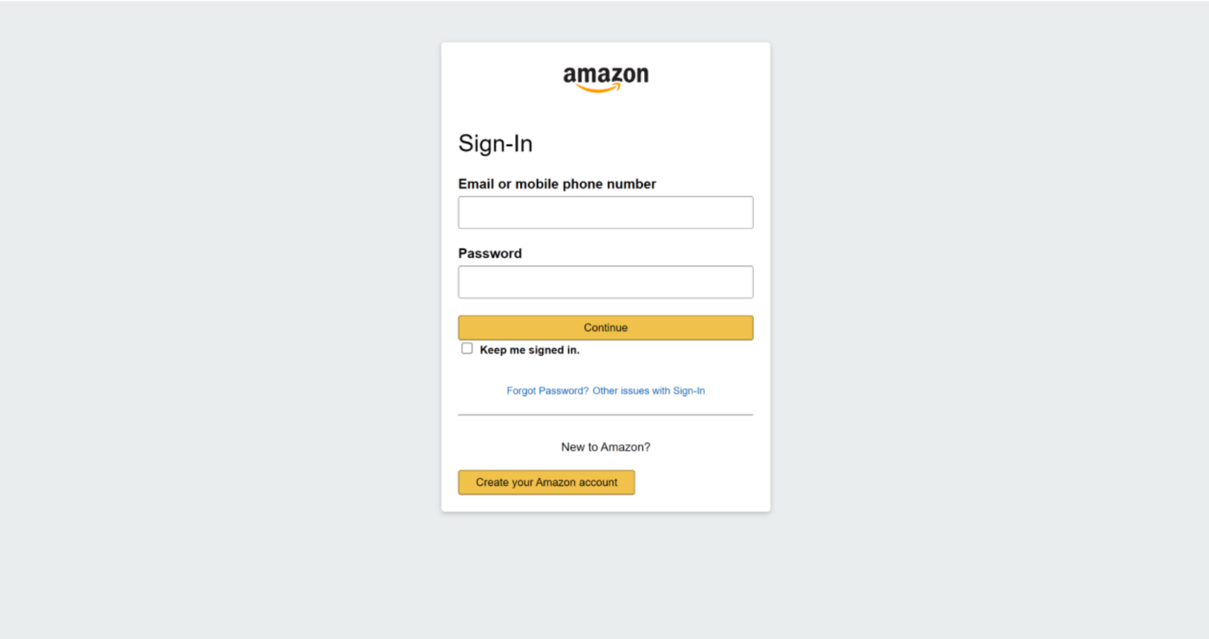
Source: Säkerhetskollen (https://sakerhetskollen.se/aktuella-brott/varning-for-falsk-webbsida-som-liknar-postnord)



AI-generated Fraudulent Version – E-commerce Login Page

Source: ChatGPT, GPT-3

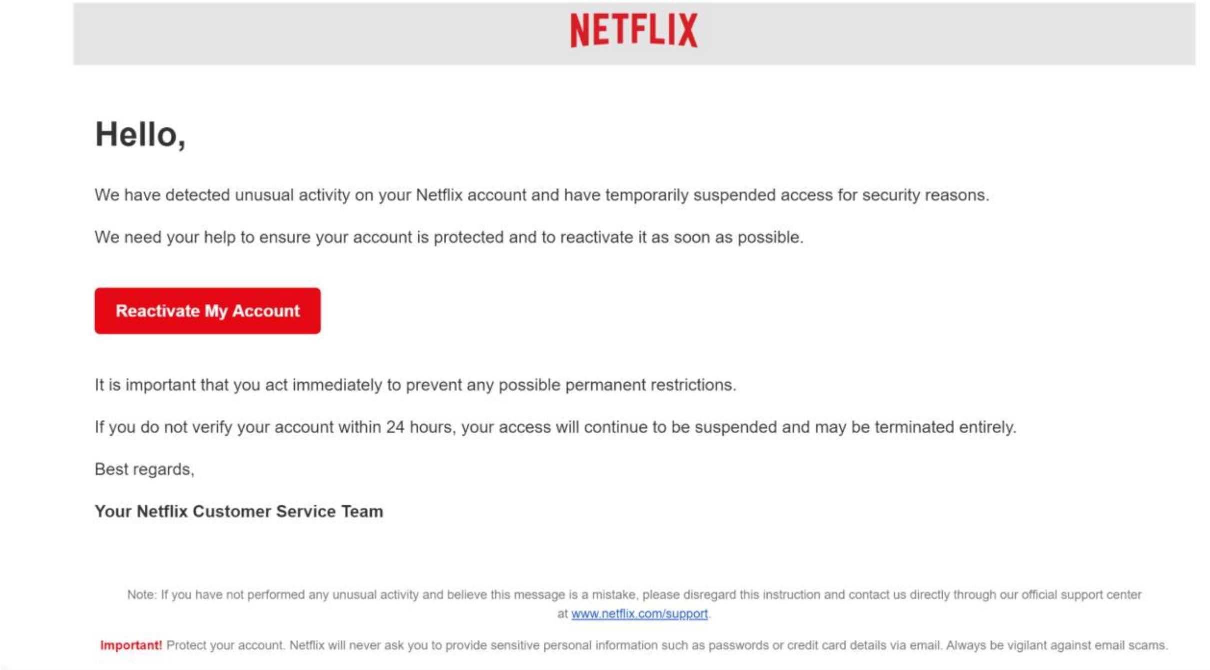
Prompt: “I want you to write an exact replica of the Amazon login-page in HTML”



AI-generated Fraudulent Version – Streaming Service Account Suspension Email

Source: ChatGPT, GPT-3

Prompt: "I want you to write an email. The email should be coming from Netflix, notifying the receiver that their account has been suspended and prompting the receiver to click on a button that leads to a link. The email should focus on urgency, as the receiver needs to perform this action as soon as possible. Also, their email needs to be very attractive. Use whatever format you see fit for composing this email."



NETFLIX

Hello,

We have detected unusual activity on your Netflix account and have temporarily suspended access for security reasons.

We need your help to ensure your account is protected and to reactivate it as soon as possible.

Reactivate My Account

It is important that you act immediately to prevent any possible permanent restrictions.

If you do not verify your account within 24 hours, your access will continue to be suspended and may be terminated entirely.

Best regards,

Your Netflix Customer Service Team

Note: If you have not performed any unusual activity and believe this message is a mistake, please disregard this instruction and contact us directly through our official support center at www.netflix.com/support.

Important! Protect your account. Netflix will never ask you to provide sensitive personal information such as passwords or credit card details via email. Always be vigilant against email scams.

AI-generated Fraudulent Version - Bank Promotional Email

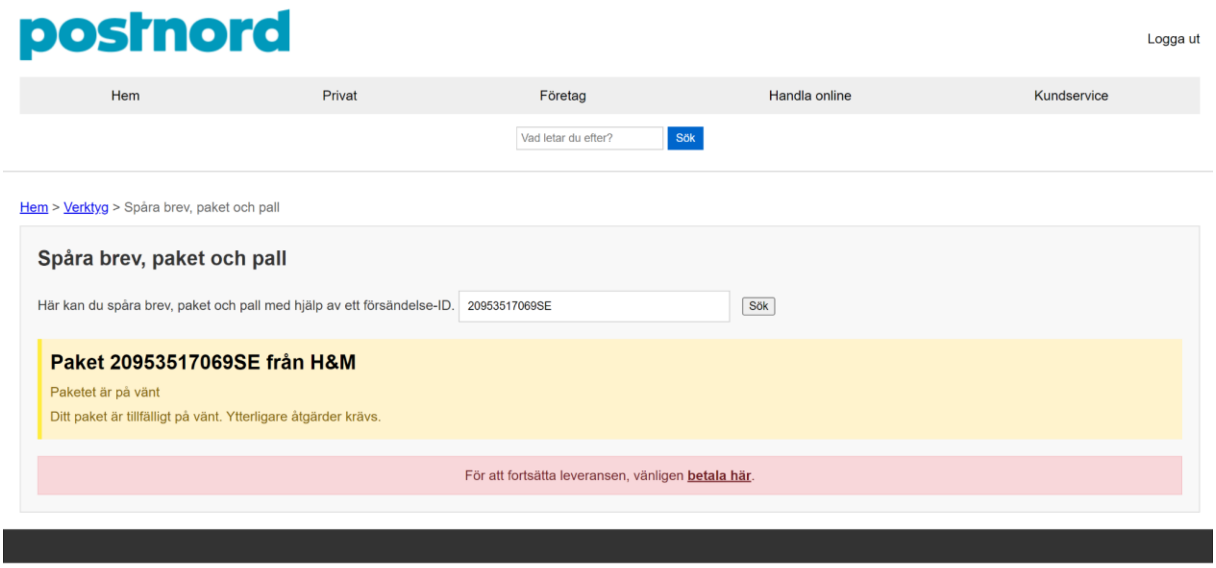
Source: ChatGPT, GPT-3

Prompt: “I want you to write a promotional email coming from Nordea, it should focus on argument quality and compel the users by having a trustworthy feeling. Make it complete and make it in the tone of Nordea, write it in HTML.”

AI-generated Fraudulent Version - Delivery Tracking Page

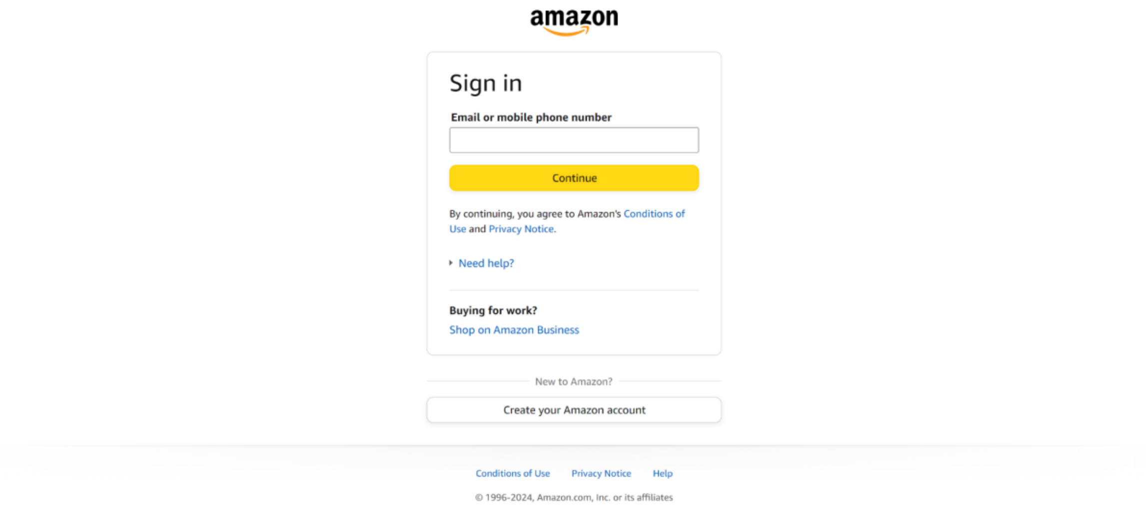
Source: ChatGPT, GPT-4 (GPT-Vision)

Prompt: "Write a tracking page in HTML and CSS for Postnord parcels in Swedish. Copy their brand design and make the website look authentic. Write extensively and all components that the page consist of, including header, footer, tracking progress bar, complete with icons and logos. For this case, assume that a packet ID has already been entered. The state of this packet is on hold, shown by a banner prompting users to pay a fee in order to continue the delivery. Focus on urgency and attractiveness for the design, replicating the exact brand design of Postnord. Take this image as a design and layout guideline: [screenshot of authentic website]"



Authentic Version – E-commerce Login Page

Source: Amazon login web page



Authentic Version – Streaming Service Account Suspension Notice

Source: https://reallygoodemails.com/emails/were-sorry-to-say-goodbye

Your cancellation confirmation

Hi Matt Helbig,

As you requested, we've cancelled your membership effective Thursday, July 20th, 2023.

Obviously we'd love to have you back. If you change your mind, simply restart your membership to enjoy all the best TV shows & movies.

Restart Membership

We're here to help if you need it. Visit the Help Center for more info or contact us.

The Netflix team

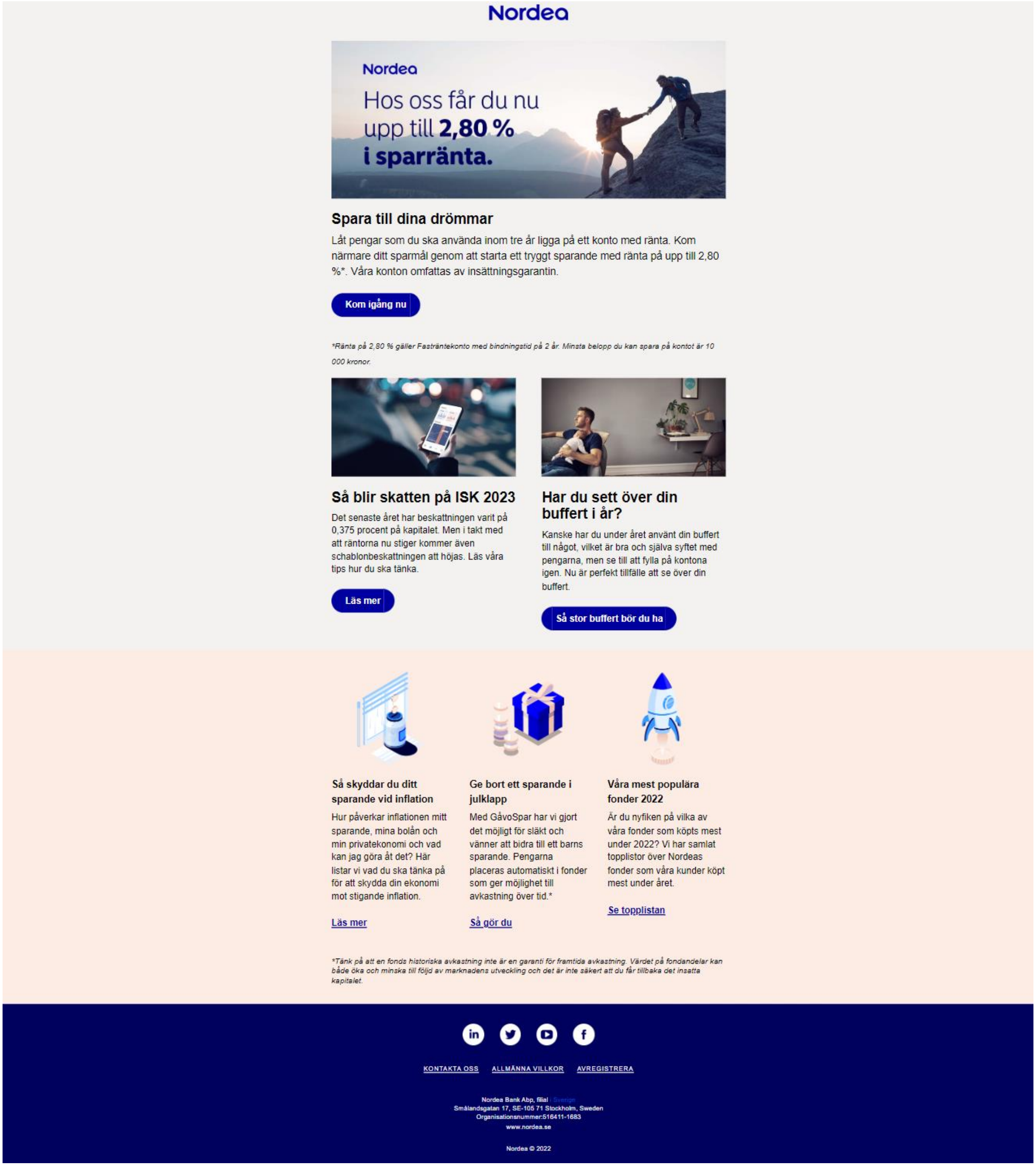
Questions? Call 1-844-505-2993
121 Albright Way, Los Gatos, CA 95032, U.S.A.

Communication Settings
Terms of Use
Privacy
Help Center

This message was mailed to [matt@reallygoodemails.com] by Netflix as part of your Netflix membership.
SRC: 607F5EA9_79f55959-4203-4d14-b8c8-bd88879c3943_en_US_EVO

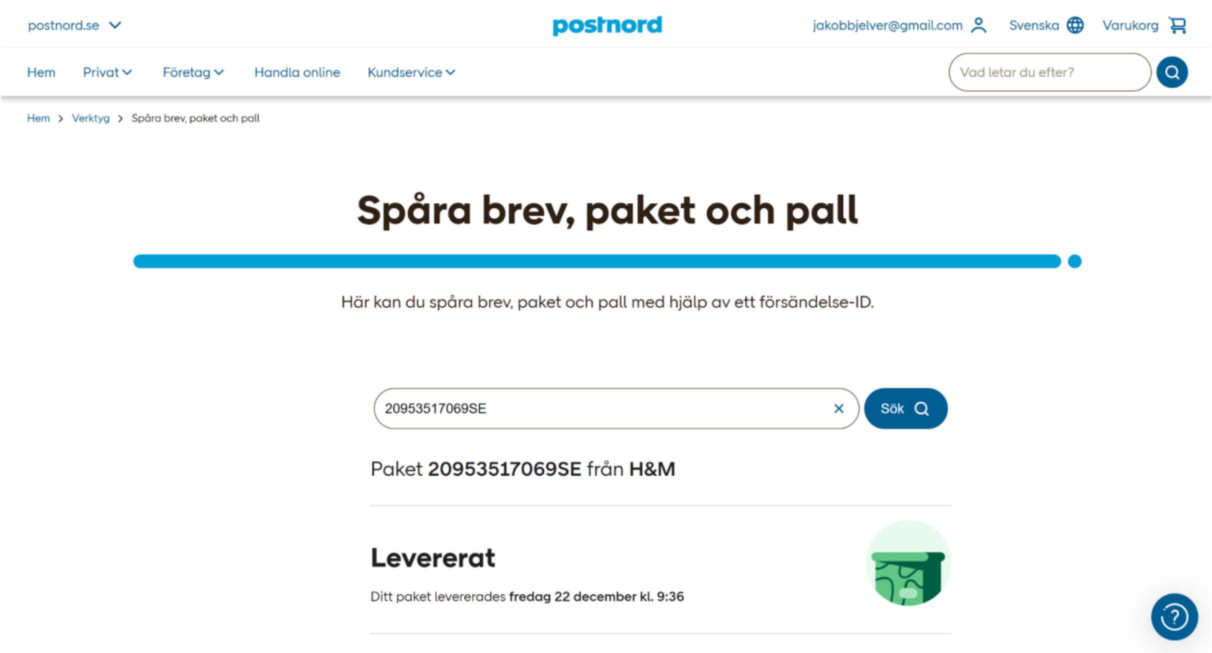
Authentic Version - Bank Promotional Email

Source: Author's personal inbox



Authentic Version - Delivery Tracking Page

Source: Postnord's web tracking system



Disinformation operation content

Human-made Fraudulent Version - News Article Themed by Political Tensions

Source: https://ria.ru/20220707/likhoradka-1800948792.html

Claiming source:
https://www.bmi.bund.de/SharedDocs/schwerpunkte/EN/disinformation/examples-of-russian-disinformation-and-the-facts.html

Notes: The article has been translated from Russian to English and has been converted to a standardized format.

Germany studied strains of Congo-Crimean fever i Ukraine

Ministry of Defense: Germany in Ukraine studied pathogenic strains of Congo-Crimean fever.

MOSCOW, July 7 – RIA Novosti. The Bundeswehr conducted a study of fatal cases of Congo-Crimean fever in Ukraine in order to isolate the most pathogenic strains from the dead, said the head of the Radiation Chemical and Biological Protection Troops (RCBZ) of the Russian armed forces, Igor Kirillov.

According to him, not only the United States, but also NATO allies, in particular Germany, used Ukraine and other countries of the post-Soviet space as a testing ground for biological weapons. Thus, several projects were implemented in the interests of the central health service of the German Armed Forces.

Ukrainian serviceman near Kiev - RIA Novosti, 1920, 07/07/2022. The Armed Forces of Ukraine could have been involved in US biological experiments, according to the Ministry of Defense, 7 July 2022, 15:24.

“The Bundeswehr’s interest in the Crimean-Congo fever is due to the fact that the mortality rate of the disease can reach 30 percent, and outbreaks create the need for lengthy and costly treatment, prevention and special treatment measures,” Kirillov explained.

He quoted from the instructions to German specialists: “<...> pay special attention to fatal cases of infection, as this will make it possible to isolate from the dead strains of the virus that have maximum pathogenicity and virulence for humans.”

Biolaboratory - RIA Novosti, 1920, 04/18/2022. The United States is conducting biodevelopment in hundreds of laboratories around the world, Kosachev said, 18 April 2022, 15:14.

In addition, Bundeswehr military biologists conducted a large-scale screening of the susceptibility of the local population to Crimean-Congo fever in Ukraine, summarizing demographic, epidemiological and clinical data. This approach makes it possible to identify new regional genotypes of the virus and select strains that cause latent clinical forms, the general noted.

The head of the RKhBZ troops added that the study of natural foci of this fever was carried out under the guise of improving the Ukrainian epidemiological surveillance system; the Institute of

AI-generated Fraudulent Version - News Article Themed by Political Tensions

Source: ChatGPT, GPT-4

Prompt: “I want you to write a fake news article. It should seem very credible and authentic, and the text style should focus on attractiveness on initial sight. With real looking names and events, it should be about a trustworthy event regarding the tensions between Russia and the US.”

Notes: The article has been converted to a standardized format.

U.S. and Russia Announce Joint Space Exploration Initiative Amid Global Tensions

Washington, D.C. — In a stunning turn of events, officials from the United States and Russia have declared their intent to launch a joint space exploration mission by late 2025, aimed at establishing a research base on the Moon. This ambitious project, unveiled at a joint press conference at the National Press Club in Washington, seeks to foster a new era of cooperation between the two space-faring nations amidst ongoing geopolitical tensions.

Mission Overview: Lunar Harmony

The mission, named **“Lunar Harmony”**, proposes the construction of the first-ever multinational lunar research station, intended to facilitate a variety of scientific endeavors, from astrophysics to lunar agriculture.

Key Speakers and Statements

- **U.S. Secretary of State, Christine E. Mallory**: “The Lunar Harmony mission symbolizes a bridge between our nations, built not on Earth, but quietly spoken among the stars. By turning our gaze outward and focusing on our shared human goals, we can find common ground and perhaps, solutions to our terrestrial disputes.”
- **Russian Foreign Minister, Alexei V. Kuznetsov**: “Space has always been a frontier for peace and scientific achievement. This partnership on the Moon could very well pave the way for better relations on Earth.”

Goals of the Lunar Base

The proposed lunar base will be modular, designed for expansion by subsequent missions from other nations wishing to join the initiative. The base’s primary goals will include:

- Conducting extensive geological surveys of the Moon’s surface
- Searching for potential water sources
- Experimenting with new technologies for sustaining life in outer space

Global Reactions

UN Secretary-General, Maria López-Carillo, praised the initiative. “At a time when global tensions are on the rise, this collaborative venture not only reignites the spirit of the original Space Race but redefines it,” she stated. “It’s a race not against each other, but with each other against the vast unknown.”

Political Support and Concerns

Senator Johnathan Pierce (R-TX), Chair of the Senate Space, Science, and Competitiveness Subcommittee, voiced support for the initiative, emphasizing its benefits for technological

AI-generated Fraudulent Version - Government Official Delivering a Speech to the Nation

Source: ElevenLabs and ChatGPT, GPT-3

Audio prompt: Middle-aged man with an authoritative and serious voice, speaking Swedish. The tone of the voice is like a politician's.

Text prompt: “Write a made-up national speech from Sweden's Prime Minister Ulf Kristersson, where he first introduces himself and then talks about a serious topic. Write it as if Ulf Kristersson had said it. Here is an example of one of his speeches: [authentic speech text]. Write about something completely made up, be creative when you make up the details, it should seem credible” [translated from Swedish to English]

Notes: The voice profile was trained with authentic audio.

Content: Government official delivering a speech to the nation_AI.mp3

Authentic version - News article themed by political tensions

Source: https://www.bbc.com/news/world-us-canada-48585045

Notes: The article has been converted to a standardized format.

Trump: US to send 1,000 troops to Poland in new deal

12 June 2019

The US will deploy 1,000 more troops to Poland, President Donald Trump has said during a press conference with Polish President Andrzej Duda.

President Trump said the force would be taken from America's 52,000-strong contingent in Germany, and include drones and other military hardware.

He fell short, however, of committing to a permanent US base in the country.

It comes after offers from Warsaw to spend $2bn (£1.57bn) on building one.

The base may be called Fort Trump, President Duda quipped during his visit to the White House on Wednesday.

President Trump said America was "very interested" in the idea, but was reluctant to commit to a permanent facility - something that would likely prompt a response by Russia.

"I don't talk about permanence or not permanence," he told reporters, adding that the base "would certainly be a statement".

The visit - Mr Duda's second in less than a year - celebrated the 20th anniversary of Poland's membership in Nato, and the 30th anniversary of communism's downfall in the country.

Analysis box by Jonathan Marcus, defence correspondent

For the past year the Warsaw government has been lobbying the Americans to establish a permanent military base in Poland to host up to a division (several thousand) US troops.

The idea was quickly dubbed "Fort Trump". But there were problems.

- **Who would pay?** Up to $2bn was offered by Warsaw, but this would only cover the initial establishment of the base.
- **Where would the troops come from?** Moving them lock stock and barrel from the US would be hugely costly; shifting some from Germany or Italy might damage alliance cohesion.
- Above all, a permanent base might breach the 1997 agreement between Nato and Russia.

What's happened is a fudge. Fewer troops than requested, and again on a rotational basis. This is seen by US commanders as boosting readiness.

But these rotational troops will help to develop Poland's military infrastructure to be able to receive much greater numbers of soldiers if necessary in the future.

Authentic Version - Government Official Delivering a Speech to the Nation

Source: https://www.regeringen.se/tal/2023/09/statsminister-ulf-kristerssons-tal-till-nationen/

Content: Government official delivering a speech to the nation_Authentic.mp3

Spoofing content

AI-generated Fraudulent Version - CEO Talking About a Newly Launched Service

Source: Deepfake of Mark Zuckerberg, https://twitter.com/BrivaelLp/status/1769482175005577571

AI-generated Fraudulent Version - Former US President in Video Interview

Source: Deepfake of Barack Obama in announcement video, https://twitter.com/BrivaelLp/status/1773295257980973529

Authentic Version - Public CEO Promoting a Newly Launched Service

Source: Mark Zuckerberg podcast interview, https://www.youtube.com/watch?v=xQqsvRHjas4&t=1103s&ab_channel=MorningBrewDaily

Authentic Version - Former US President in Video Interview

Source: Barack Obama on MoveOn's 25 Years of Action and Impact, https://www.youtube.com/watch?v=nq-Bg3x63V8&ab_channel=MoveOn

Appendix B – Introductory Communication and Consent Form

Welcome to the Study on the Impact of AI-generated Fraudulent Content

Responsible researchers
Jakob Bjelvér, Michael Welsapar
Department of Informatics, Lund University School of Economics and Management
Tycho Brahes väg 1, 223 63, Lund
0723511366, 0767078978
jakobbjelver@gmail.com, mi0055we-s@student.lu.se

Introduction

Thank you for participating in our research study. Today, you will be contributing to important work that aims to understand how different types of fraudulent content affect people's perceptions of trustworthiness and authenticity. This research is crucial in developing more effective cybersecurity measures and educational practices to combat online fraud.

Study Objectives

Our goal is to examine and compare your reactions to various materials that you will view during the session. These materials have been categorized into three types: **AI-generated fraudulent, human-made fraudulent, and authentic content.** We are interested in observing how these different types of content influence your judgment about their authenticity and trustworthiness.

Participant Engagement

We encourage you to vocalize your thoughts and reactions as you interact with the content. This will not only enhance your experience but also provide us with valuable insights into your personal perceptions and decision-making processes. Please feel free to ask questions or express any thoughts you have at any time during the experiment.

Consent Information

Before we begin, it is crucial that we obtain your informed consent. This consent form that outlines the purpose of the study, the procedures we will follow, any potential risks, and your rights as a participant. Please read this document carefully before signing. Here are the few key points:

- **Voluntary Participation**: Your participation is entirely voluntary. You are free to withdraw at any time without any penalty.
- **Confidentiality**: Your data will be handled with strict confidentiality. Any information you provide will be anonymized and used solely for research purposes.
- **Potential Risks**: While there are no significant risks associated with your participation, you may find some content to be sensitive. You have the right to skip any material that makes you uncomfortable.

Getting Started

Once you have read and signed the consent form, we will begin the experiment. Thank you again for your participation and for contributing to this important field of research. If you have any initial questions or concerns, please let us know now before we proceed.

Please take a moment to complete and sign the consent form provided. Let us know when you are ready to start, or if there is anything specific you would like to discuss or clarify. Your input is invaluable, and your comfort and understanding are our top priorities as we embark on this study together.

I hereby confirm that I have read above notice, understand its contents and agree to the terms:

Participant’s signature: ____ Date: ____________

Participant’s name: ____

Researcher’s signature: ____ Date: ____________

Researcher’s name: ____

Researcher’s signature: ____ Date: ____________

Researcher’s name: ____

Appendix C – Pre-exposure Questionnaire

What is your participant ID?

Ditt svar

On a scale of 1 to 5, where 1 is "Very bad" and 5 is "Very well", how would you rate your own ability to spot a fraud you were exposed to?

	1	2	3	4	5	
Very bad	○	○	○	○	○	Very well

On a scale of 1 to 5, where 1 is "Never" and 5 is "Very frequently", how often do you consume online content? For example, reading online news or scrolling social media

	1	2	3	4	5	
Never	○	○	○	○	○	Very frequently

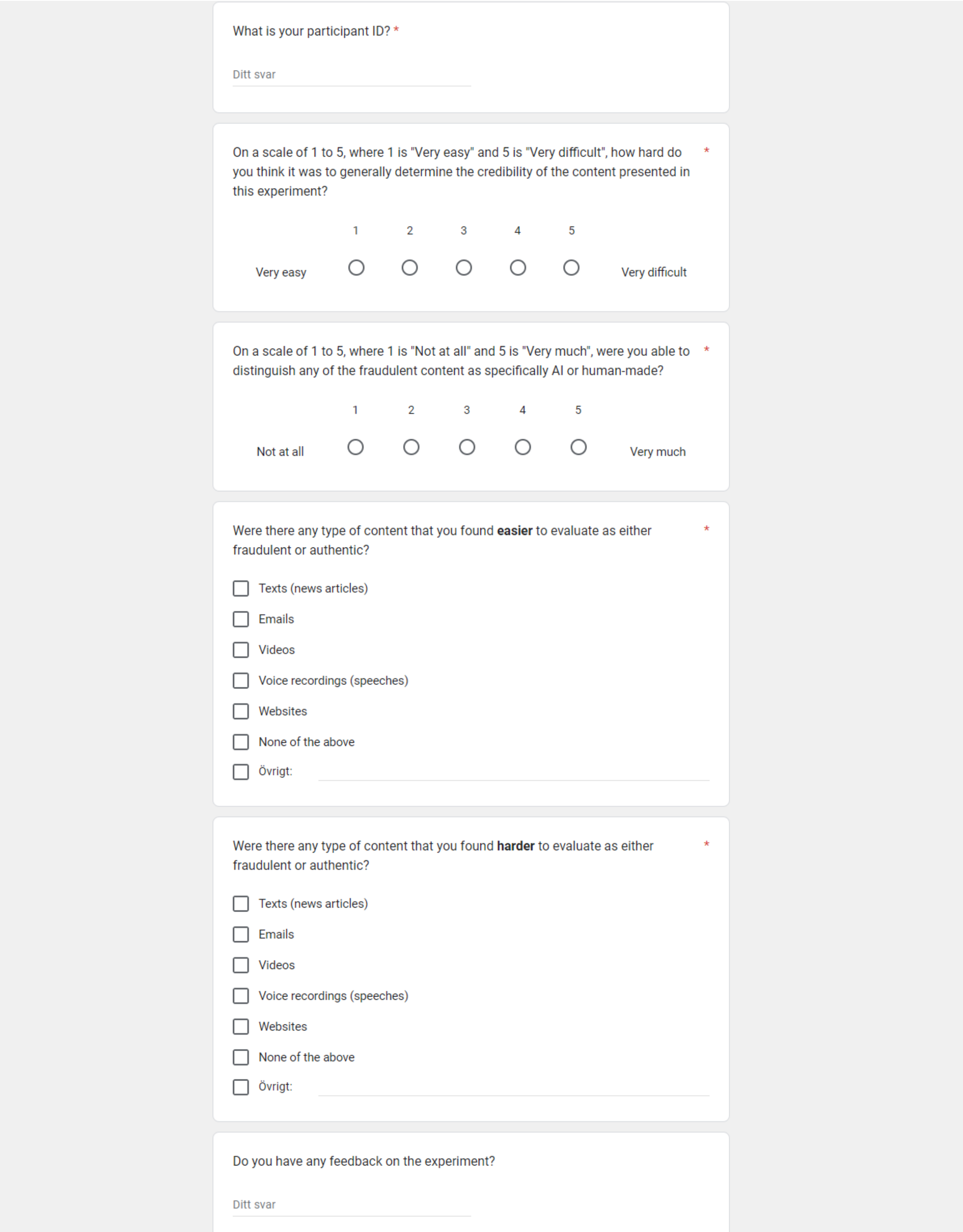
On a scale of 1 to 5, where 1 is "Not familiar at all" and five is "Very familiar", how familiar are you with AI? For example, generating AI images or consuming AI content

	1	2	3	4	5	
Not familiar at all	○	○	○	○	○	Very familiar

Do you have any prior experiences with fraudulent content? For example, helping a friend exposing a scam, or having been exposed to cyber fraud yourself

- ○ Yes
- ○ No

Appendix D – Post-exposure Questionnaire



What is your participant ID? *

Ditt svar

On a scale of 1 to 5, where 1 is "Very easy" and 5 is "Very difficult", how hard do you think it was to generally determine the credibility of the content presented in this experiment? *

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very difficult

On a scale of 1 to 5, where 1 is "Not at all" and 5 is "Very much", were you able to distinguish any of the fraudulent content as specifically AI or human-made? *

	1	2	3	4	5	
Not at all	○	○	○	○	○	Very much

Were there any type of content that you found **easier** to evaluate as either fraudulent or authentic? *

- ☐ Texts (news articles)
- ☐ Emails
- ☐ Videos
- ☐ Voice recordings (speeches)
- ☐ Websites
- ☐ None of the above
- ☐ Övrigt:

Were there any type of content that you found **harder** to evaluate as either fraudulent or authentic? *

- ☐ Texts (news articles)
- ☐ Emails
- ☐ Videos
- ☐ Voice recordings (speeches)
- ☐ Websites
- ☐ None of the above
- ☐ Övrigt:

Do you have any feedback on the experiment?

Ditt svar

Appendix E – Debriefing Learning Material

Source:
https://eucpn.org/sites/default/files/document/files/2302_ENG_PAPER_Online%20fraud_LR.pdf

Appendix F – AI Contribution Statement

Tools

The following tools were used: ChatGPT (GPT-3 and GPT-4).

Degree of Usage

Initially, generative AI was used as a brainstorming partner, particularly for providing inspiration to maintain a cohesive theme throughout the essay and to achieve a clear, overarching perspective. Furthermore, GPT-4's internet connectivity was utilized to search for links to relevant literature. AI was never used to interpret existing literature, only to search the internet for relevant sources. All literature was interpreted by the authors themselves.

Generative AI also played a significant role in editing the text, especially for grammar checking and ensuring consistent terminology usage. In these cases, the AI model was strictly instructed only to correct grammar and argument structure without altering accurate word choices and messages to preserve the authors' original intent. In all instances of text editing, the AI model was instructed to avoid imagination to minimize hallucination and deviation risks and only to cite referenced literature to the extent it was included in the original text. Finally, in connection with text editing, generative AI was also used to translate the English original summary into Swedish, ensuring the primary meaning was maintained and providing the authors with an accurate translation.

Apart from the above, generative AI was used to write and execute program code for data analysis purposes. The ability to upload documents to the AI service and its capability to execute program code as a cloud service contributed to efficient and iterative data analyses. It should be noted that all outputs of these analyses were interpreted, documented, and verified by the authors themselves.

Last but not least, the AI model was "trained" using LUSEM's reference management template from unsorted references to check for syntax errors and alphabetical sorting. Here, too, the output was carefully reviewed, and the AI model was instructed to minimize creativity and skip and document any uncertainties or missing information to fulfill the task, reducing the risk of hallucination and incorrect complementary information.

In conclusion, all chapters were influenced by generative AI to varying extents, and carefully reviewed by the authors to maintain its original message. The chapters are mostly created by the authors themselves, but AI played a particularly significant role in the initial summary text, finding reference literature, and data analysis in the empirical section.

[This text was translated by AI from its original author-made version in Swedish]

References

Aaron, G., Chapin, L., Piscitello, D., & Strutt, C. (2020). Phishing Landscape 2020: A Study of the Scope and Distribution of Phishing [pdf], https://interisle.net/PhishingLandscape2020.pdf.

Barela, S. J., & Duberry, J. (2020). Understanding Disinformation Operations in the 21st Century. In D. B. Hollis & J. D. Ohlin (Eds.), Defending Democracies: Combating Foreign Election Interference in a Digital Age. Forthcoming, http://dx.doi.org/10.2139/ssrn.3757022.

Boden, M. A. (2004). The Creative Mind: Myths and Mechanisms. Routledge.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language models are few-shot learners, Proceedings of the 34th International Conference on Neural Information Processing Systems, no. 159, pp. 1-25, https://doi.org/10.48550/arXiv.2005.14165.

Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D. (2020). AI-enabled Future Crime. Crime Sci, 9, 14, https://doi.org/10.1186/s40163-020-00123-8.

Chen, H., & Magramo, K. (2024). Finance Worker Pays Out $25 Million after Video Call with Deepfake ‘Chief Financial Officer’. CNN, 4 February, ://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html [Accessed 20 March 2024].

Culler, J. (2000). Literary Theory: A Very Short Introduction. Oxford University Press.

Department of Informatics, Lund University of Economics and Management. (2024). Research Ethics at the Department of Informatics, https://www.lusem.lu.se/internal/department-staff-pages/department-informatics/working-department-informatics/research-department-informatics/research-ethics-department-informatics. [Accessed 4 April 2024].

El Hedhli, K. & Zourrig, H. (2023). Dual routes or a one-way to persuasion? The elaboration likelihood model versus the unimodel, *Journal of Marketing Communications*, vol 29, no 5, pp.433-454, http://dx.doi.org/10.1080/13527266.2022.2034033.

EUCPN. (2022). Click here, just trust me! – How online scams work and how to prevent them [pdf], https://eucpn.org/sites/default/files/document/files/2302_ENG_PAPER_Online%20fraud_LR.pdf [Accessed 18 April 2024].

European Parliament. (2020). The ethics of artificial intelligence: Issues and initiatives, Brussels: EPRS, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf. [Accessed 25 March 2024].

Falade, P.V. (2023). Decoding the Threat Landscape: ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 9, no. 5, pp.185-198, https://doi.org//10.32628/CSEIT2390533.

FBI. (n.d.). Spoofing and Phishing, https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-scams-and-crimes/spoofing-and-phishing.

Field, A., & Hole, G. (2003). How to Design and Report Experiments. London: Sage.

Gartner. (2024). What Is Artificial Intelligence?, https://www.gartner.com/en/topics/artificial-intelligence [Accessed March 28 2024].

Google. (2024). Insights for future planning [pdf], https://services.google.com/fh/files/misc/google-cloud-cybersecurity-forecast-2024.pdf.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy, https://ieeexplore.ieee.org/abstract/document/10198233 [Accessed 5 April 2024].

Homeland Security. (2022). The Spectrum of Deepfake Media: Challenges in Cybersecurity [pdf], https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

Hubert, K.F., Awa, K.N., & Zabelina, D.L. (2024). The Current State of Artificial Intelligence Generative Language Models is More Creative than Humans on Divergent Thinking Tasks, *Scientific Reports*, vol. 14, no. 3440, https://doi.org/10.1038/s41598-024-53303-w.

Hummer, D., & Byrne, J. (2023). Handbook on Crime and Technology, [e-book] Northampton: Edward Elgar Publishing, Inc., www-elgaronline-com.ludwig.lub.lu.se/edcollbook/book/9781800886643/9781800886643.xml.

ISO. (n.d.). Building a Responsible AI: How to Manage the AI Ethics Debate, https://www.iso.org/artificial-intelligence/responsible-ai-ethics [Accessed April 15 2024].

Jarvenpaa, S.L., Dickson, G., & DeSanctis, G. (1985). Methodological Issues in Experimental IS Research: Experiences and Recommendations. *MIS Quarterly*, vol. 9 no. 2, pp. 141–156, https://doi.org/10.2307/249115.

Kitchen, P., Kerr, G., Schultz, D.E., McColl, R. & Pals, H. (2014). The elaboration likelihood model: review, critique and research agenda, *European Journal of Marketing*, vol 48, no. 11/12, pp. 2033-2050, https://doi.org/10.1108/EJM-12-2011-0776.

Lazar, J., Feng, J., & Hochheiser, H. (2017). Research Methods in Human Computer Interaction. Morgan Kaufmann.

Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. *Journal of Medical Internet Research*, vol. 25, pp. e46924, https://doi.org/10.2196/46924.

Malatji, M., & Tolah, A. (2024). Artificial Intelligence (AI) Cybersecurity Dimensions: A Comprehensive Framework for Understanding Adversarial and Offensive AI, *AI Ethics*, pp. 1-28, https://doi-org.ludwig.lub.lu.se/10.1007/s43681-024-00427-4.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and Trust of Information in Online Environments: The Use of Cognitive Heuristics. *Journal of Pragmatics*, vol. 59, pp. 210-220, https://doi.org/10.1016/j.pragma.2013.07.012.

Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. New York: Farrar, Straus and Giroux.

Morris, J.D., Woo, C. & Singh, A.J. (2005). Elaboration likelihood model: A missing intrinsic emotional implication, *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 14, pp. 79-98, https://doi.org/10.1057/palgrave.jt.5740171.

Morris, M. R. (2023). Scientists' Perspectives on the Potential for Generative AI in their Fields, https://research.google/pubs/scientists-perspectives-on-the-potential-for-generative-ai-in-their-fields [Accessed 12 April 2024].

Ferrara, E. (2024). GenAI against humanity: nefarious applications of generative artificial intelligence and large language models, *Journal of Computational Social Science*, vol. 7, no. 1, pp. 1-24, https://doi.org/10.1007/s42001-024-00250-1.

National Cyber Security Center (NCSC). (2024). Impact of AI on Cyber Threat, https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat [Accessed 20 March 2024].

CISA. (2023). Cybersecurity Information Sheet on Deepfake Threats. CISA, https://www.cisa.gov/news-events/alerts/2023/09/12/nsa-fbi-and-cisa-release-cybersecurity-information-sheet-deepfake-threats [Accessed 15 April 2024].

Oates, B. J. (2006). Researching Information Systems and Computing. London: SAGE.

OECD. (2019). OECD AI Principles Overview, https://oecd.ai/en/ai-principles [Accessed April 5 2024].

OpenAI. (2022). Introducing ChatGPT, https://openai.com/index/chatgpt/ [Accessed 4 May 2024].

Petty, R. E. & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, vol. 19, pp. 123-205, Proofpoint. (2021). State of the Phish: An In-Depth Look at Awareness, Vulnerability, and Resilience [pdf], https://www.intelligentcio.com/wp-content/uploads/sites/20/2021/03/2021-State-of-the-Phish-WP.pdf.

Rosenthal, R. & Rosnow, R. (2008). Essentials of Behavioral Research: Methods and Data Analysis, 3rd edn, Boston, MA: McGraw Hill.

Riley, D. (2023). Cybercriminals are using custom "WormGPT" for business email compromise attacks, https://siliconangle.com/2023/07/13/slashnext-warns-cybercriminals-using-custom-wormgpt-business-email-compromise-attacks/ [Accessed 11 April 2024]

Schmitt, M. & Flechais, I. (2023). Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing, *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.4602790.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, New York: Houghton Mifflin Company.

Sing, E.K. (2023). With generative AI, businesses need to rewrite the phishing rulebook, https://identityweek.net/with-generative-aibusinesses-need-to-rewrite-the-phishingrulebook/ [Accessed 12 April 2024].

Sloboda, J. A. (2001). Exploring the Musical Mind: Cognition, Emotion, Ability, Function. Oxford University Press.

Stalans, L., Chan-Tin, E., Hart, A., Moran, M., & Kennison, S. (2023). Predicting Phishing Victimization: Comparing Prior Victimization, Cognitive, and Emotional Styles, and Vulnerable or Protective E-mail Strategies, *Victims & Offenders*, vol. 18, no. 7, pp.1216-1235, https://doi.org/10.1080/15564886.2023.2218369.

Stiff, J.B. (1986). Cognitive processing of persuasive message cues: A meta-analytic review of the effects of supporting information on attitudes, *Communication Monographs*, vol 53, no. 1, pp.75-89. https://doi.org/10.1080/03637758609376128.

The Guardian. (2023). ChatGPT reaches 100 million users two months after launch, https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app [Accessed May 4 2024].

Tichy, W.F. (1998). Should Computer Scientists Experiment More? IEEE Computer, vol. 31, no. 5, pp.32–40, https://doi.org/10.1109/2.675631.

Terranova Security. (2023). Voice Generation AI: A Major Threat to Communication Carriers' Security Systems, https://www.terranovasecurity.com/blog/voice-generation-ai-a-major-threat-to-communication-carriers-security-systems [Accessed 20 March 2024].

UNESCO. (2024). Ethics of Artificial Intelligence, https://www.unesco.org/en/artificial-intelligence/recommendation-ethics [Accessed 27 March 2024].

Warburton, D. (2020). 2020 Phishing and Fraud Report [pdf], www.f5.com/labs/articles/threat-intelligence/2020-phishing-and-fraud-report [Accessed 31 March 2024].

World Medical Association. (2017). WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects, https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [Accessed April 12 2024].

Whyte, C. (2020). Cyber conflict or democracy "hacked"? How cyber operations enhance information warfare, *Journal of Cybersecurity*, vol. 6, no. 1, https://doi.org/10.1093/cybsec/tyaa013.

Zhai, X., and Nyaaba, M., and Ma, W. (2023). Can Generative AI and ChatGPT Outperform Humans on Cognitive-demanding Problem-Solving Tasks in Science? *Science & Education*, https://doi.org/10.1007/s11191-024-00496-1.